eudeco

Modelling the European data economy

WP1 Mapping the scene

D1.2 Report on the analysis of framework conditions

Final version – 31 August 2015

Funded by the Horizon 2020 Framework Programme of the European Union

D1.2 Report on the analysis of framework conditions

Work package	WP1 Mapping the scene
Lead author	Daniel Bachlechner (FRAUNHOFER)
Contributing authors	Sven Rung (FRAUNHOFER), Michael Friedewald (FRAUNHOFER), Antonio Alfaro (ROOTER), Laura Merlo (ROOTER), Franz Ruz (ROOTER), Sven Abels (ASCORA), Marcel Gabor (ASCORA), Leonardo Haut (ASCORA), Anthony Dupont (SIGMA), Ida Laude (SIGMA), Stephanie Dos Santos (SIGMA), Bart Custers (LEIDEN), Helena Ursic (LEIDEN)
Due date	M6 (July 2015)
Date	31 August 2015
Version	1.0

Type	Report
Dissemination level	Public

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 645244. This document does reflect the authors view only. The European Commission is not responsible for any use that may be made of the information the document contains.

Document history

Version	Date	Author(s)	Notes
0.1	27 April 2015	Daniel Bachlechner (FRAUNHOFER)	Initial proposal for the deliverable's structure as basis for discussion
0.2	26 June 2015	Daniel Bachlechner (FRAUNHOFER)	Integration of first drafts of sections 3 and 5, and detailing of the section 4 structure
0.3	30 July 2015	Daniel Bachlechner (FRAUNHOFER)	Revision of section 1; integrations of contributions to sections 2 to 5
0.4	11 August 2015	Daniel Bachlechner (FRAUNHOFER)	Minor corrections and addition of missing parts
0.5	18 August 2015	Daniel Bachlechner (FRAUNHOFER)	Minor corrections
0.6	25 August 2015	Daniel Bachlechner (FRAUNHOFER)	Revision based on SIGMA's review comments
0.7	27 August 2015	Daniel Bachlechner (FRAUNHOFER)	Minor corrections
1.0	31 August 2015	Daniel Bachlechner (FRAUNHOFER)	Final check

Disclaimer

Table of contents

List of figures

List of abbreviations

APEC	Asian Pacific Economic Cooperation
API	Application programming interface
B2B	Business-to-Business
BCR	Binding corporate rules
BDaaS	Big Data as a Service
BDVA	Big Data Value Association
CAS	Complex adaptive system
CC	Creative Commons
CC0	Creative Commons Zero
CDS	Critical Data Studies
CJEU	Court of Justice of the EU
CSV	Comma separated values
CPU	Central processing unit
DO	Data obfuscation
DOS	Denial of service

DPA	Data protection authority
EC	European Commission
EC2	Amazon Elastic Compute Cloud
ECHR	European Convention on Human Rights
EDI	Electronic data interchange
EDPS	European Data Protection Supervisor
EMR	Elastic MapReduce
ERP	Enterprise Resource Planning
EU	European Union
HDFS	Hadoop Distributed File System
HTML	Hypertext Markup Language
HTTPS	Hypertext Transfer Protocol Secure
IaaS	Infrastructure as a Service
ICT	Information and communication technology
IETF	Internet Engineering Task Force
IoE	Internet of Everything
IP	Intellectual property
IPR	Intellectual property right
ISMS	Information Security Management System
ISO	International Organization for Standardization
IT	Information technology
JSON	JavaScript Object Notation
KML	Keyhole Markup Language
LAPSI	The European Thematic Network on Legal Aspects of PSI
MS	Member State
NIS	Network and Information Security
OAuth	Open Authorization
OECD	Organisation for Economic Co-operation and Development

OWL	Web Ontology Language
PPP	Public-Private Partnership
PSI	Public sector information
RAID	Redundant Array of Independent Disks
RDF	Resource Description Framework
RDMS	Relational Database Management System
REST	Representational State Transfer
RFC	Request for Comments
SaaS	Software as a Service
SLA	Service-level agreement
SME	Small and medium-sized enterprises
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
TCP	Transmission Control Protocol
TRIPS	Agreement on Trade-Related Aspects of Intellectual Property Rights
UDP	User Datagram Protocol
URL	Uniform Resource Locator
UK	United Kingdom
US	United States
W3C	World Wide Web Consortium
WTO	World Trade Organization
XML	Extensible Markup Language
XSS	Cross-Site Scripting

Executive summary

D1.2 reports on the findings of an analysis of framework conditions relevant in the context of the data economy from a legal, a socio-economic and a technological perspective. The analysis is a key foundation for the creation of an initial, heuristic model of the European data economy. The deliverable clearly pays most attention to the discussion of the framework conditions but also presents a first integrative discussion as well as a pragmatic approach towards a conceptual framework.

The conceptual framework will be revised and extended within the scope of the creation of the model. For now, a definition of data reuse is provided as well as an initial overview of actor groups and data types relevant in the context of the data economy. With respect to the understanding of data reuse, data recycling (i.e., using data several times for the same purpose), data repurposing (i.e., using data for different purposes than for which they were initially collected) and data recontextualisation (i.e., using data in another context than in which they were initially collected) are differentiated. Data repurposing and data recontextualisation are considered to be particularly interesting with respect to the European data economy. This is taken into account within the scope of the perspective-specific discussions of framework conditions.

Legal perspective: The analysis of legal framework conditions focused on selected areas of public and private law. With respect to public law, data protection law, privacy law, non-discrimination law, other human rights, data retention law, data localisation law, freedom of information and open data – regulation of public sector information (PSI), cybersecurity law and national security law were investigated. Intellectual property rights (IPRs) law, contractual law, completion law and consumer protection law were at the centre of attention with respect to private law.

Key findings with respect to the legal analysis are:

- Laws and regulations can be both barriers to and enablers to data reuse. However, the analysis has showed that most often their influence is negative, i.e. laws and regulations act as barriers to data reuse.
- It can be argued that data protection law, privacy law and anti-discrimination law are the public law areas that have the greatest impact on data reuse.
- On the private law side, Intellectual property law is the most exposed field.
- Although the current literature about the legal framework for data reuse only focuses on a limited scope of laws, it is critical to bear in mind the legal landscape for data reuse is very complex and often remote legal rules can be applicable.
- A general data law seems to be attractive at first sight but may be difficult to implement and enforce.

Socio-economic perspective: The analysis of socio-economic framework conditions focuses on the societal perspective first and takes afterwards and economic perspective. With respect to the societal perspective, perceptions of big data issues, the psychology of mass self-communication, strategies for limiting the analysis of personal data and aspects of the creation of a data culture are addressed. With respect to the economic perspective, the financing of economic activity, the institutional landscape, research and development, the labour market and characteristics of the data market are addressed.

Key findings with respect to the socio-economic analysis are:

- Creating a data culture is probably a key prerequisite for enabling the extensive development of a data economy.
- Many individuals don't have the necessary knowledge to estimate the consequences of the digital traces they leave behind.
- Research on the data market, so far, has mainly focussed on open data in general and PSI in particular.
- The European Union (EU) recognizes big data as a driver of the future European economy and urges that its potential be exploited.
- The growing venture capital funding gap between the United States (US) and Europe leads to a location disadvantage for Europe.

Technological perspective: The analysis of technological framework conditions focuses on data collection and delivery, scalability and data management, data transmission and exchange formats, data aggregation, analysis and interpretation, reliability and availability, technological responsibility and last but not least security and privacy. With respect to each of the focus areas challenges and barriers as well as approaches and example cases are discussed.

Key findings with respect to the technological analysis are:

- Many organizations rely on *de facto* standards that are created and maintained by a single actor in the market.
- The instability of interfaces, the evolution of data formats and legacy systems are key challenges associated with data collection and delivery.
- Data transmission is of prime importance in a data economy that is characterised by a large number of actors interacting with each other.
- Innovative data analysis approaches and algorithms are often critical for the success of data reuse in new contexts.
- Transferring responsibility for technology leads to severe challenges related to security and privacy.

1 Introduction

This section details the purpose and scope of D1.2, its structure and relationship to other deliverables as well as the methodology applied.

1.1 Purpose and scope

The main purpose of D1.2 is to provide an overview of framework conditions relevant in the context of the European data economy and their impact. The set of framework conditions described will be extended and revised through the course of the project. Just like D1.1 and D1.3, D1.2 provides a central pillar with respect to the initial heuristic model of the European data economy. The heuristic model will be described in D2.1.

Whenever possible, the relevance of the framework conditions was investigated for different groups of actors. Similarly, framework conditions in Europe were compared with those in other regions of the world, particularly with those in the US. Additionally, differences with respect to framework conditions between EU Member States (MS) have been elaborated, whenever relevant ones were detected. Finally, efforts were made to anticipate future developments with respect to framework conditions relevant for the European data economy.

1.2 Structure of the document

Section 2 provides an outline of concepts particularly relevant in the context of the European data economy. This section is the basis for the development of a conceptual framework. Its focus is restricted to the most relevant actor groups and data types. Throughout the course of the project further concepts will be added and the descriptions of the existing concepts will be refined. The concepts are selected having the notions of complex adaptive systems (CASs) in mind.

Sections 3 to 5 focus on framework conditions relevant in the context of the European data economy from a legal, a socio-economic and a technological perspective, respectively. In section 3, particular attention is paid to the legal framework observing both private and public law rules. Section 4 looks at the European data economy as well as related challenges from a societal and an economic perspective. Aspects of the technological infrastructure as well as characteristics of data are what section 5 focuses on.

Section 6 sketches the results of a first attempt to bring the legal, the socio-economic and the technological perspective together. Moreover, it is highlighted, with respect to the heuristic model, what is already there and what still has to be developed. Section 7 concludes the deliverable.

1.3 Relationships to other deliverables

D1.2 provides, together with D1.1 and D1.3, the foundation for the development of an initial heuristic model of the European data economy. The heuristic model will be described in D2.1. The refined model

described in D3.1 as well as the final model described in D4.1 build upon the heuristic model. Apart from the model and its iterations, the analysis of framework condition is relevant for the recommendations given as well as the observatory developed during the last phase of the EuDEco project. Without doubt, D1.2 can be considered a key deliverable of the EuDEco project. Nevertheless, the insight provided should not be considered final. Many points may be revised and extended several times until the end of the project.

1.4 Methodology

In terms of methodology, the framework conditions were identified by means of a review of relevant literature. The literature review was systematic, which means that it was tried to get a profound understanding of the breadth of the academic discourse. However, as framework conditions touch a large number of aspects, it was not possible to go into details with respect to all aspects. Aspects which are considered particularly relevant for EuDEco will thus be further investigated in later phases of the project. Whenever possible, the search strategy and the criteria for inclusion and exclusion of literature were documented.

Interim results with respect to relevant framework conditions were presented at the initial cluster workshop and discussed with the UEG members. The discussion focused on the relevance as well as the completeness of the set of framework conditions that had already been identified at that point in time. Moreover, UEG members were asked to assess the impact of specific framework conditions in Europe on the emerging data economy. In addition to the initial cluster workshop, a EuDEco workshop was held at the 28th Beld eConference. The design of this workshop was similar to the one of the initial cluster workshop. At both occasions, interim results related to legal, socio-economic and technological framework conditions were presented and discussed with external experts and other interested individuals.

The results of the literature review as well as the feedback collected during the workshops are discussed within the scope of D1.2. The analysis and discussion focuses on both the framework conditions from the three perspectives and the relations between legal, socio-economic and technological framework conditions.

In order to allow an integrative discussion of the three perspectives as well as to facilitate the creation of the heuristic model of the European data economy, the three perspective-specific analyses of framework conditions refer, as far as possible, to common sets of actor groups and data types. Therefore taxonomies of actors and data were created in parallel with the perspective-specific research on framework conditions. The two taxonomies were both inspired by existing taxonomies.

2 Conceptual framework

2.1 Understanding of data reuse

Before it is possible to map the scene regarding data reuse it is important to consider first what exactly is data reuse. The term data reuse in its broadest sense suggests that there is initial (primary) use of data and subsequent (secondary) use of data, the reuse of data. The distinction between use and reuse may imply different aspects however. Take the following example.

An insurance company collects patient data in order to have a proper client database used for billing the insurance premiums that are due and to reimburse medicines, treatments and therapies. When they use a client's address for sending them a bill, they will do this monthly, quarterly or annually. In that sense, they periodically reuse the address more than once for the same purpose. This is a form of data reuse or data recycling.

In case the same insurance company starts using the data to assess risks of patients in order to determine risk-based insurance premiums (e.g., higher premiums for people at risk or showing unhealthy behaviour like smoking, not exercising, etc. and lower premiums for people at low risks showing healthy behaviour), they are reusing the data for a different purpose. This is also a form of data reuse; that can be classified as data repurposing.[1]

When the same health insurance company starts selling the data, other companies may also make use of the data, for instance, for marketing their products to particular target groups. The data is then reused in a (sometimes completely) different context. This may cause issues of contextual integrity or ethical issues, since data may have a different meaning or may be interpreted differently in another context (Nissenbaum, 2004). For instance, health data may be interpreted differently by a physician than by a health insurance company. This is also a form of data reuse, which can be classified as data re-contextualisation.

In summary, we can thus distinguish data reuse as:

- Data recycling – using data several times for the same purpose
- Data repurposing – using data for different purposes than for which they were initially collected
- Data recontextualisation – using data in another context than in which they were initially collected

In this deliverable, we will mainly focus on the second and third meaning of data reuse, as they are expected to be of most added value in the European data economy. The added value of big data and

[1] Loshin (2011) makes the distinction between data reuse and data repurposing.

open data is likely to increase in large, aggregated datasets in which data from different sources (e.g., from different social sectors and industries) are combined.

Note that from a legal perspective the different forms of data reuse as described above are categorized slightly differently. This is because in the EU data protection regulation the collection of data is also regarded as a form of data processing.[2] As a result, it may be suggested that data processing always starts with data collection and that subsequent actions like data storage, preparation and analyses are all next steps in data processing. This is, however, not the way in which data reuse is considered in the EU directive on the protection of personal data. The directive rather focuses on data repurposing, as stated in Article 6.1(b): personal data must be "collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes". For data repurposing typically (additional) informed consent of the data subjects has to be obtained.

In section 3, the EU regulations that prohibit data repurposing and data recontextualisation are further discussed. These forms of data reuse are usually referred to as *function creep* and are not allowed unless there is a legal basis for it. Such a legal basis may exist, for instance, when people provide (additional) informed consent. A problem here is that it is sometimes unknown for which other purposes collected data may prove to be useful a few years later (Custers, 2004, p. 215). Sometimes new tools for data analyses may also yield new insights like novel patterns and relations in datasets. Not collecting data that can easily be collected is increasingly regarded as a waste of economic resources (Bygrave, 2002; Zittrain, 2000). For these reasons, many corporations collecting personal data formulate the purposes of their data collection very broadly, so that concrete purposes do not necessarily have to be known at the time of collection.[3] As a result, there appear to be differences in the legal interpretation of data repurposing and the more common understanding of it by data subjects. From a legal perspective data repurposing or function creep only exists in case the secondary use of personal data goes beyond the purposes specified in advance. Since these are in general formulated very broadly, this may rarely be the case. From a more common understanding, however, people may consider data repurposing to be the case when personal data is used for purposes they did not expect their data to be used for (Custers, van der Hof, Simone, & Schermer, 2014; European Commission, 2011b).

[2] Article 2 (b) of EU Directive 95/46/EC on the protection of personal data defines processing of personal data as: "any operation […] performed upon personal data […] such as collection, recording, organisation, storage [etc.]"
[3] For instance, IT company Cisco states at http://www.cisco.com/web/siteassets/legal/privacy.html: "We collect personal information for a variety of reasons, such as processing your order, providing you with a newsletter subscription, personalizing your experience, or in connection with a job application. We will inform you of the purpose for collecting personal information when we collect it from you and keep it to fulfil the purposes for which it was collected or as required by applicable laws or for legitimate purposes. We may combine the information we collect from you with information obtained from other sources to help us improve its overall accuracy and completeness, and to help us better tailor our interactions with you. We may also collect information relating to your use of our websites through the use of various technologies, including cookies."

Literature on expectations of data subjects on how their data is used and what they consider acceptable shows that there is a considerable gap between the practices of data controllers and the expectations of data subjects. For instance, US-based research has shown that race and ethnicity play an influential role in how people use social media and share personal information (Correa & Jeong, 2011). Users show concern for privacy, although there seems to be an incongruity between public opinion and public behaviour: people tend to express concern about privacy, but when asked about it, they routinely disclose personal information because of convenience, discounts, and other incentives, or a lack of understanding of the consequences (2013, p. 49; Regan, 2003). These tensions between attitudes and practices were also found by Acquisti and Gross (2006). A possible explanation for this tension may be that users do not connect the disclosure of their data at a particular time with the use of their data later, for different purposes, by different organisations and in a different context. Such connections may not be transparent when there are long periods of time between the data collection and actions based upon the processing or sharing of such information and when data are sold to other companies or between the (primary) data use and the (secondary) data reuse. Also the ways in which data are processed may not be transparent. For instance, when the information collected is used for profiling, such profiling techniques, by their nature, tend not to be visible processes for data subjects (Bygrave, 2002; Custers, 2004).

As will be discussed in section 4, from a societal perspective it may in some situations be desirable to reuse data for different purposes or in different context and other situations public support may be lacking. From an economic perspective it may be very valuable to combine data from different sectors, but it may also occur that companies prefer to refrain from data sharing, for instance, when it concerns their business models or intellectual property (IP).

From a technological perspective (section 5), a minimum condition for data reuse is an adequate technological infrastructure. Obviously, standardised approaches in information technology (IT) architecture and data formats further facilitate data reuse and data exchanges, but also aspects like scalability, aggregation, reliability, availability and security are relevant in this respect.

2.2 Actor groups

This section contains a first collection of actor groups relevant in the context of the data economy. This list will be extended and revised during the development of the EuDEco model of the data economy. With respect to some of the groups, concrete actors will be named. Moreover, relationships among the actor groups and actors will be described in forthcoming deliverables.

With respect to most actor groups, concrete actors can be persons, companies or other entities.

Among the actor groups referenced in this document are:

- regulators
- technology providers
- service providers
- data users
- data collectors, data aggregators, etc.
- data owners
- data subjects
- data processors
- data controller

Regulators are public or private bodies that set the regulatory framework (rules). With respect to data reuse, among them are, for instance, data protection authorities, competition regulators and consumer protection bodies.

Technology providers develop, produce and sell software applications and/or hardware for use by their clients in the context of the data economy.

Service providers provide specific services to their clients. In the context of the data economy services may be related to, for instance, data collection, data aggregation, data management, data analysis or data visualisation.

Data users exploit data to their advantage.

All activities along the value chain may be conducted by different actors or by one actor. Consequently, one actor can be data collector, data aggregator, etc. at the same time. The actor can also be the data user. Alternatively, an actor can outsource selected activities to service providers.

Data owners can usually authorize or deny access to certain data. They are not necessarily owners of the data in legal terms.

The terms *data subject*, *data processor* and *data controller* are mostly used in the context of the EU data protection law.

Data subjects are identified or identifiable persons whom particular personal data is about (Information Commissioner's Office, n.d.). The definition of the subject is not limited to natural persons and MSs are free to broaden the scope of their national provisions to include legal entities. The European Court of human rights has not yet ruled on the question, whether data attached to a specific legal entity may also be considered personal (Kuner, 2007, p. 91).

Data processors are the natural or legal person, public authority, agency or any other body, which processes personal data on behalf of a data controller (Article 2, General Directive).

Data controller are the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data (Article 2, General Directive).

2.3 Data types

This section contains a first collection of types of data relevant in the context of the data economy. This list will be extended and revised during the development of the model of the data economy. Moreover, relationships among the data types will be described in forthcoming deliverables.

big data has intentionally not been included in the list of types of data as big data does not describe a specific type of data. While the term may seem to reference the volume of data, that isn't always the case. The term big data may also refer to the technology that is required to handle large amounts of data.

Among the types of data referenced in this document are:

- raw data
- open data
- government data
- academic data
- personal data
- structured data
- unstructured data

Raw data is a term for data that has not been subjected to processing or any other manipulation. An alternative term is *primary data*.

Open data is accessible public data that people, companies, and organisations can use to launch new ventures, analyse patterns and trends, make data-driven decisions and solve complex problems. All definitions of open data include two basic features: the data must be publicly available for anyone to use, and it must be licensed in a way that allows for its reuse. Open data should also be relatively easy to use, although there are gradations of *openness* (Gurin, 2014).

Government data can relate to publicly accessible or to undisclosed sensitive governmental information (e.g. sensitive information related to the national security). An alternative term is PSI. According to the European Thematic Network on Legal Aspects of PSI (LAPSI) glossary, PSI is defined as the wide range of information that public sector bodies collect, produce, reproduce and disseminate in many areas of activity while accomplishing their institutional tasks. PSI may include social, economic, geographical,

cadastral, weather, tourist, and business information.[4]. PSI acquires a specific legal meaning within the EU, since it has been provided with a minimum set of rules contained in the Directive 2003/98/EC of 17 November 2003 on the re-use of PSI.

Academic data relates to the data needed to reproduce investigations as well as the outputs of research. While the former may be spreadsheet data, videos, code and field-specific proprietary formats, the later may be posters, presentations and other traditional outputs including research papers and theses. The heterogeneity in data between academic fields is a fundamental difference between academic and government data (Hahnel, 2015).

Both government data and academic data are increasingly made available as open data.

Personal data is any information relating to an identified or identifiable natural person.[5] A special regard should be given to sensitive data, e.g. data that relates to a person's racial or ethnic origin, his political or religious opinion and his health and sexual life.

Structured data is data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets. Structured data first depends on creating a data model.

Unstructured data is data that can't be so readily classified and fit into a box: photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

The differentiation between structured and unstructured data is particularly relevant when the data economy is looked at from technological perspective. From a technological perspective, the data format is also used to differentiate between types of data. Examples are JavaScript Object Notation (JSON) data, eCl@ss data or tabular data.

From a socio-economic perspective, it is relevant to differentiate types of data based on their purpose or source. With respect to the former, there is, for instance, training data, with respect to the later, social media data, sensor data and experimental data. Additionally, the differentiation between surface data and deep data is relevant from this perspective. Surface data, according to Manovich (2011) is analysed through statistical, mathematical or computational techniques. Relevant methods in the context of deep data are hermeneutics, participant observation, thick description, semiotics and close reading.

[4] http://www.lapsi-project.eu/lapsifiles/lapsi_glossary.pdf

[5] An identifiable person is anyone who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity (Article 2 (a) of the General Directive; also see Article 29 Working Party (2007, p. 4).

There are numerous data types which denote what the respective data is about. In health care, a field were sensitive data plays an important role, for instance, patient data, genetic data, health data and medical data are relevant. With respect to the advertising industry, a field where data reuse is highly relevant, consumer data, sales data and price data are of particular relevance. Further examples are traffic data, financial market data, telemetry data and satellite image data.

3 Framework conditions: A legal perspective

3.1 Introduction

Section 3 of the deliverable D1.2 examines the legal framework for data reuse. The aim of this analysis is to describe the EU legal landscape in which data reuse is taking place and to consider all relevant legal issues that might impact data reuse activities.

The analysis focuses on framework conditions in the EU. However, it also addresses the differences between the EU and some other frameworks around the world. Due to a large amount of commercial activities (including data reuse) between the EU and the US, the American regulation is of particular interest.[6] In addition, the analysis takes into account the differences between the legal frameworks of the EU MSs, since not all legal areas have been unified or harmonised.[7] As the law in this area has been continuously changing, this analysis also anticipates on future developments of the framework conditions relevant for the European data economy, such as the proposal for General Data Protection Regulation[8]. In subsequent stages of the project a number of carefully selected case studies will be used to further concretize this general, introductory overview. In this later deliverable sector-specific and industry-specific legal hurdles, such as telecommunications law, financial law or other areas where big data is used, will be discussed in more detail.[9]

The EuDEco project, which this analysis is a part of, addresses the complexity of the data economy and the data reuse through the CAS theory. By adopting the CAS approach the project aims to get a better understanding of the relations in the complex data value network[10] and their interrelations to the different relevant frameworks (legal, economic, societal, and technological). The analysis at hand has no ambition to address these interactions in more detail, as this is a discussion that should be raised in further stages of the EuDEco project. However, it does not turn a blind eye on the fact that understanding the law as one of the cornerstones of the EU data economy system is critical in order to

[6] See for example section 3.2.4.1.7.
[7] See for example section 3.2.4.5.
[8] Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), COM (2012) 11 final, 25.1.2012.
[9] See also section 3.3.2.
[10] As opposed to the data value chain a data value network is a complex rather a corruptive than a linear setting.

illustrate the complexity of the research subject.[11] Namely, the legal framework is interdependent with the broader economic, technical and political environment, influenced by multiple (inter)national law enforcement agencies, governmental bodies and private organisations. Through the prism of CAS theory, this environment would be perceived a global CAS. This global setting has a corruptive rather than linear logic. Legal framework has to adapt to its unforeseen evolution, led by the interplay of governmental and private actors trying to regulate the area, the overlap of national approaches including the jurisdictional dilemmas and strong pressure from the business side to take the bottom-up approach to decision-making.

The analysis of the legal framework is structured as follows. Section 3.2 examines the legal framework for data reuse in the EU. Sections 3.2.1 to 3.2.3 explain our approach to legal analysis. Secondly, sections 3.2.4 and 3.2.5 list relevant legal areas in public law and private law and point out the rules that influence data reuse. Section 3.3 addresses a number of considerations that do not relate directly to the legal framework but are important to consider before the framework is integrated into the model.[12] Section 3.4 concludes.

3.2 Legal framework

3.2.1 Positioning data reuse within legal borders

Reuse of data raises numerous legal questions, which can be analysed from different legal perspectives. Data protection law, IP and competition regulation are clearly in the front line, but, as will be shown, also other areas of law may be relevant to some extent.

In order to structure the analysis two realms are distinguished, private law and public law, following the traditional dichotomy of laws (Hage & Akkermans, 2014, p. 38). Law that regulates the vertical relationship between the state and private parties shall be deemed public whereas law that applies to horizontal dealings among private parties shall be labelled private (Rosenfeld, 2013, p. 126). Public law encompasses constitutional law, administrative law and international law. Administrative law is understood in a broader sense, as the collection of rules that regulate interactions between government agents and civilians or private organisations. Hence, social security law, tax law, environmental law and data protection law[13] can all be listed under the umbrella of administrative law (Hage & Akkermans,

[11] Law is an important feature in the data economy, both on the EU as well as on the international level. Some authors have found some significant CAS features for the data protection law subject matter, and claim it is very likely that the data (protection) law itself could also be described as a CAS Zhang and Schmidt, Aernout H. J. (2015).
[12] This is especially true since we agreed to follow the CAS approach, for which a mere analysis of the legal conditions does not suffice. See page 3 for more details.
[13] Data protection law straddles the boundaries between public and private law, criminal and civil law. This makes it difficult to firmly place data protection law within any one of the legal categories traditionally employed by the doctrines of private international law Bygrave (2000). Given the stronger human rights foundation that data protection law enjoys in the EU, it is intentionally classified under the realm of public law.

2014, p. 38). Among all possible areas of public law this analysis only takes into consideration those that are *prima facie* relevant for data reuse activities. Private law, on the other hand, regulates market transactions between citizens and focuses in the areas such as IP, contractual, competition and labour law (Hage & Akkermans, 2014, p. 38). Again, the analysis is only focused on the legal branches that prove relevance to data reuse activities.

The selection of the areas of public law is limited to:

- data protection law
- privacy law
- non-discrimination law
- other human rights
- data retention law
- data localisation law
- freedom of information and open data – regulation of PSI
- cybersecurity law
- national security law

The selection of areas of private law is limited to:

- IPRs law
- contractual law (including licences)
- competition law
- consumer protection law

Also, as has been already raised in the literature (Custers, Calders, Schermer, & Zarsky, 2013, p. 61), the laws in the area of data reuse often overlap. For example, data protection laws and anti-discrimination laws can be both applicable to a particular state of facts. While this overlap can actually add to better legal protection (Custers et al., 2013, p. 81), it also increases the complexity of the legal framework. Furthermore, the importance of (big) data on the market has revived the link between data protection and competition law. The growing collection, processing and use of consumer transaction data for commercial ends are proving an increasingly important source of competitive advantage, which could be an increasing source of consumer detriment (Currie, 2013). This analysis endeavours to identify such overlaps and explore how they should be addressed most adequately.

3.2.2 Big data and open data

At the outset, it should be stressed that the emergence of big data has amounted to the complexity of the discussion on data reuse, since the current regulation may not sufficiently respond to the challenges

related to big datasets (see for example Rubinstein 2013)[14]. Furthermore, there have been claims that our society should expect a substantial loss of benefits of big data, if it attempts to confine it within an obsolete legal framework (Information Commissioner's Office, 2014). These reactions call for a comprehensive analysis that also takes into account the challenges of big data and recent policy responses on the EU level.

Another relevant phenomenon is open data. The idea that more data should be available for free has been strongly advocated in the public sector, in the EU most visibly by the Directive 2003/98/EC of the European Parliament and of the Council on the reuse of PSI [2003] OJ L 345 (hereafter Directive on PSI). Due to many positive side effects such as transparency and trust, open data initiatives have been slowly moving towards the private sector. Admittedly, open data as a concept is unlikely to be seen attractive for the private businesses (Reggi, 2011). Still, opening up at least a minimum amount of private data could result in great social benefits and by giving the business sufficient incentives it should be indeed possible to achieve greater openness in the private sector as well. Hence, the legal analysis should address the developments on both sides, in the public as well as the private sector.

[14] For further discussion also see Tene and Polonetsky (2013) and Kuner, Cate, Millard, and Svantesson (2012).

3.2.3 The legal landscape for data reuse

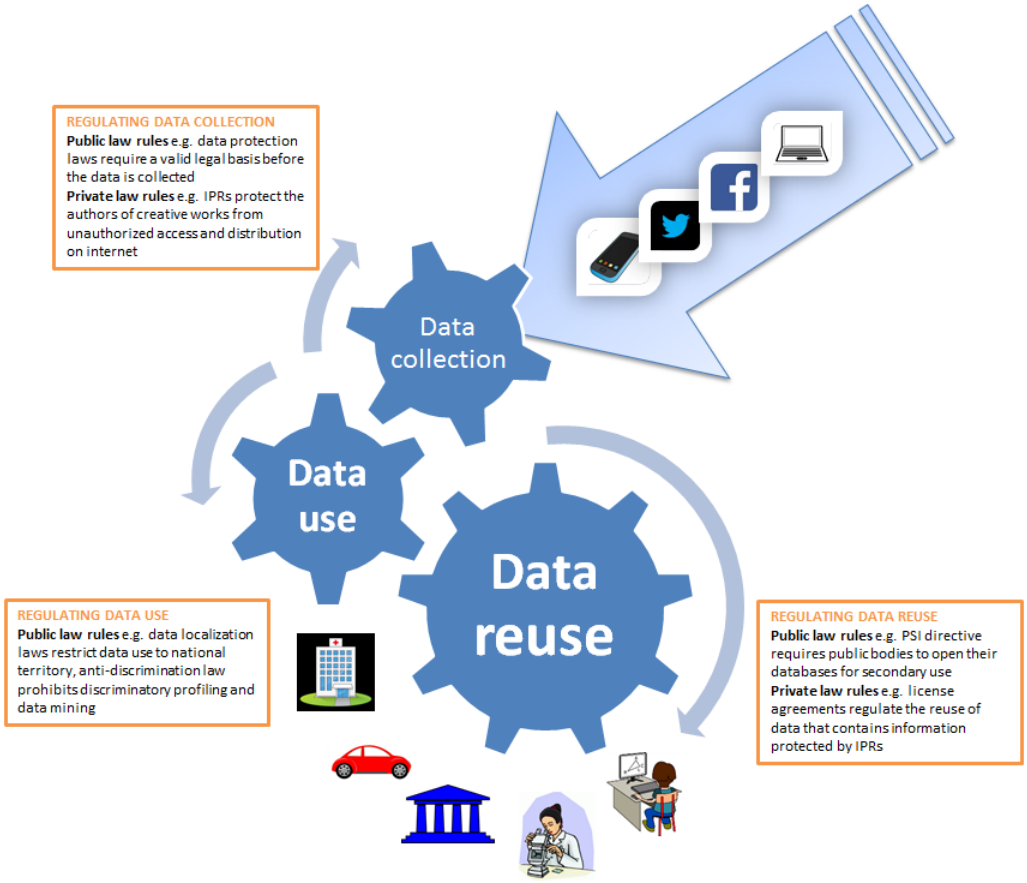

Figure 1: The legal landscape of data reuse

Figure 1 is a high level illustration of the legal landscape for data reuse. The upper cogwheel presents activities related to data collection including the channels through which the data is collected and transferred to data users. The cogwheel in the middle presents data usage and the cogwheel in the bottom presents its re-use. While these concepts are useful to illustrate business reality, they do not entirely correspond to the legal framework, which operates with less tangible terms such as data processing and data repurposing. This simple conceptualisation only serves as a basis of the legal analysis and will be elaborated in later stages of the projects.

Furthermore Figure 1 depicts the rules that pertain to different phases of the data (reuse) lifecycle. Orange boxes present the regulatory landscape for each phase. The first one relates to data collectors, such as governments, hospitals and cloud services providers. Data protection law only allows collection under strict conditions. For instance, medical data can only be obtained if a data subject agrees with that and has to be adequately protected. The second group of rules relates to data users, such as public bodies and privately held companies. There are pieces of information that are prohibited to be used or

they can only be used under strict conditions. For instance, the information that is protected by copyrights can only be used if the creator allows it. In addition, some data can only be used within one country and it may be prohibited to send it over the national borders. The third group of the rules applies to data reusers e.g. research institutions, advertisers and pharmaceutical companies. Before these subjects are able to reuse data, they need to consider various legal restrictions. For instance, if the dataset they plan to reuse contains information protected by copyright, they have to comply with the restrictions of the licence agreement issued by the IPRs holder.

3.2.4 Public law

Section 3.2.4 deals with public law rules that regulate data reuse. Every sub-section begins with a table of applicable EU laws and continues with a general introduction to the chosen legal area. Sub-sections end with an analysis of the applicable rules through the prism of data reuse.

3.2.4.1 Data protection rules

EU DATA PROTECTION LAW	
Primary EU law	European Convention on Human Rights (Art. 8) Charter of the fundamental rights of the EU (Arts. 7 and 8) Treaty on the functioning of the EU (Art 16)
Secondary EU law	General Data Protection Directive Regulation concerning the protection of individuals with regard to the processing of personal data by Community institutions and bodies E-privacy Directive General Data Protection Regulation (draft proposal)

The modern economy is based on data collection and its flows.[15] Since a considerable part of this information relates to individual persons, personal data protection rules set some significant limits to data reuse. It is thus sensible to start the legal analysis with an overview of data.

3.2.4.1.1 Introduction

Data protection laws date back to 1980s when the Organisation for Economic Co-operation and Development (OECD) adopted the first set of globally accepted and recognized personal data protection

[15] This also follows from the European Commission's statements: "Digital content that can serve the needs of every European, and generate new opportunities for business. It's the life that flourishes in a vibrant digital ecosystem." Kroes (2013)

rules, also known as Convention 108 (OECD, 1980). Over the following years, numerous states worldwide followed its example and, based on the original OECD text, adopted their own data protection laws (European Union Agency for Fundamental Rights & Council of Europe, 2014, pp. 16–17). Clearly, there have been some important differences in the approaches, particularly as regards the level of restrictiveness. For example, Spain, Portugal, Germany and Italy are known by their very limiting rules, while India, China and Brazil have followed a more relaxed approach (DLA Piper, 2015). In its early attempts to establish the EU single market, the EU realized that data could only be freely transferred across the Union if data protection laws were harmonised (European Union Agency for Fundamental Rights & Council of Europe, 2014, p. 17). In 1995, the European parliament and of the Council adopted the Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L 281 (hereafter General Directive) which has remained the fundamental European legal act in the area of data protection until today.

3.2.4.1.2 Definitions

In order to understand the scope and the impact of the EU data protection law it is first necessary to define the main concepts, such as the object and the subject of legal protection.

The object that data protection laws protect is personal data. According to Article 2 of the General Directive the definition of personal data includes any information relating to an identified or identifiable natural person.[16] A special regard should be given to sensitive data, e.g. data that relates to a person's racial or ethnic origin, his political or religious opinion and his health and sexual life. The General Directive prohibits processing of such data unless one of the few exceptions applies.

At this point it is important to bear in mind that anonymised data (i.e. personal data disconnected from the source) is not protected by data protection laws. Anonymisation is a method applied to personal data in order to achieve irreversible deidentification (Article 29 Working Party, 2014, p. 7). Its two main techniques are randomization and generalization (Article 29 Working Party, 2014, p. 3). Anonymisation should be differentiated from the concept of pseudonymisation (i.e. concealing the identity by using numbers instead of names). As the latter still enables (de)-identification of the initial source, the processing of pseudonymised data should remain restricted.[17]

The subject of legal protection, also referred to as *data subject*, is an identified or identifiable person whom particular data is about[18]. The definition of the subject is not limited to natural persons and MSs are free to broaden the scope of their national provisions to include legal entities. The European Court

[16] An identifiable person is anyone who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity (Article 2 (a) of the General Directive; also see Article 29 Working Party (2007, p. 4).
[17] The discussion on ananyminsation and pseudonimisation techniques goes beyond the scope of this analysis. More details can be found in Guidance on anonymization techniques Article 29 Working Party (2014).
[18] https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/

of Human Rights has not yet ruled on the question, whether data attached to a specific legal entity may also be considered personal (Kuner, 2007, p. 91).

The next category that should be defined, are data users. In principle, these could be either individuals or legal persons including its affiliates. Departments or other organizational sections of a company are excluded from the definition (European Union Agency for Fundamental Rights & Council of Europe, 2014, p. 50). In legal terms users can be split into two groups – data controllers and data processors. A data processor is a natural or legal person, public authority, agency or any other body, which processes personal data on behalf of the controller (Article 2, General Directive). A data controller is the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data.

When personal data is processed as part of data reuse activities, data protection law normally apply. The reusers are considered data controllers and they are subject to all obligations and rights of the data protection legislation (LAPSI, 2012, p. 1).

3.2.4.1.3 Privacy principles

The General Directive builds on the OECD idea of strong privacy principles as expressed in Convention 108. However, the scope of the convention has been importantly expended. The directive clarifies some undefined concepts (e.g. consent), introduces a few European-specific provisions (territorial scope) and expands data subjects' rights. This section summarizes the EU directive by guiding the reader along the seven OECD privacy principles and explaining them through the prism of the General Directive's provisions.

- The collection limitation principle states that "[t]here should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject".[19] Moreover, the use limitation principle sets forth that "[p]ersonal data should not be disclosed, made available or otherwise used for purposes other than those specified, except a) with the consent of the data subject; or b) by the authority of law". The two principles incorporate the same idea as the provisions on fair/lawful processing and on purpose limitation, both stipulated in Articles 6 and 7 of the General Data Protection Directive. The latter improves the OECD's concept of use limitation by listing six exhaustive criteria under which processing of data is allowed:
    (a) the data subject has unambiguously given his consent; or
    (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract; or

[19] This principle is sometimes referred to as the *principle of minimality*, see Bygrave (2002, p. 341).

(c) processing is necessary for compliance with a legal obligation to which the controller is subject; or
(d) processing is necessary in order to protect the vital interests of the data subject; or
(e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed; or
(f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject.

- According to LAPSI (2012), the reuse of the information, most often, will not be necessary for the performance of a contract (see (b) above), neither will it be needed to comply with a legal obligation (see (c) above), to carry out a public task in the public interest or to protect the vital interest of the data subject (see (d) above). Thus, two other possible legal basis for data reuse remain, namely consent (see (a) above) and legitimate interest (see (f) above). As getting the consent of every person of whom personal data is contained in the information will be too time-consuming and laborious, this is not a very feasible option. Therefore it is most likely that the processing will be justified through a legitimate interest provision (the so-called balancing provision - see point (f) above) with which the interest of the controller or the third party to which the data is disseminated is balanced with the interest of the data subject, especially with regard to the respect for his fundamental rights to privacy and data protection (LAPSI, 2012, p. 24). Only if another fundamental right is served by the reuse of PSI containing personal data, most commonly the right to freedom of speech, will there be a situation in which two equal interests must be balanced. This is unlikely to be the case when data is reused in commercial purposes. Therefore, a commercial reuser will often find it difficult to demonstrate a legitimate legal basis for the processing of personal data (LAPSI, 2012, p. 25). This also explains why LAPSI researchers consider complete prohibition of personal data reuse the most feasible solution to privacy dilemmas related to reuse of PSI, wherein they acknowledge that this measure would leave the economic potential of the European PSI unutilized.
- Data minimisation principle is a component of the use limitation principle. Its objective is to prevent unlimited collection and retention of personal data. Data minimisation can be seen as a serious barrier to data reuse. Especially in the era of big data, the idea of minimising the information collected in the past collides with the business intention to analyse and use as much data as possible. Data minimisation can be in a direct conflict with particular business models, in particular those that base their business on selling the collected data for further use/reuse (Rubinstein, 2013).[20]

[20] As a practical example, see for instance https://www.healthbank.ch/. This platform encourages people to share their health data, which is later sold to pharmaceuticals or some other third party for secondary use and research.

- The principle of purpose limitation prohibits the use of data whenever it goes beyond the initial purpose, e.g. the purpose a data subject consented to. In other words, whenever personal data is processed, its reuse is, in principle, prohibited. Namely, reuse will rarely be in the scope of the initial purpose, since at the moment when the data is collected it is often difficult to predict all possible ways in which the data could be exploit in the future. In the opinion 03/2013 on purpose limitation (Article 29 Working Party, 2013, p. 14), Article 29 Working Party acknowledged the current tendency for reuse of data by the private sector ('big data') and 'open data' and 'data sharing' initiatives proposed by many governments, including EU legislative initiatives. Based on these trends, it emphasized that purpose limitation principle applied not only to personal data held by the private sector but also to personal data held by the public sector. In addition, the Working Party pointed out that the principle of purpose limitation continued to apply to personal data even if such data had been made publicly available.
- Considering its restrictive nature, the principle of purpose limitation often bottlenecks data reuse. If the company decides to use the data for a purpose that differs from the one described in their terms and conditions (for example now the company wants to share and reuse the data with a third party), it might run into some important legal barriers (Rubinstein, 2013).[21] Data reuse would be only allowed if the controller would receive an additional, broader consent from the data subject, however this can be a great logistical hurdle, especially in the era of big data.
- The purpose specification principle states that "[t]he purposes for which personal data are collected should be specified and that the data may only be used for these purposes". In the EU law, these purposes must be specified and made manifest by the controller before the processing of data starts (Article 29 Working Party, 2013, p. 15). Hence, in case of reusing certain personal information this specific reuse has to be specified already at the moment of data collection. In practice, it is unlikely that all possible reuses can be defined or predicted in advance. Admittedly, data reuse can be included in the purposes specified by the data controller by using a broad purpose formulation. However, this can be seen as circumventing the purpose of legislator and processing based on it can be considered illegitimate if not illegal.
- The data quality principle states that "[p]ersonal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date". General Directive expands the wording by adding that the data should also be relevant, not excessive in relation to the purpose and kept no longer than necessary for the processing. To comply with the relevancy requirement it is necessary to omit processing of the irrelevant (unimportant, not related to the business activity) data. Data accuracy means that data has to reflect the true condition and be, if necessary, updated. Article 6 of the General Directive reads that every reasonable step must be taken to ensure that data,

[21] Facebook is an apparent example of such a business model. On behalf of Belgian data protection authority (DPA), the researchers from van Alsenoy et al. (2015) have identified a number of issues where FB's business activities clearly violate the ideas of the EU data protection laws.

which are inaccurate or incomplete, are erased or rectified. Finally, data must be kept in a form, which permits identification of data subjects for no longer than it is necessary for the purposes for which the data were collected or for which they are further processed (European Union Agency for Fundamental Rights & Council of Europe, 2014, pp. 72–75).

- Since principle of data quality ensures that the data remains accurate, it makes it more suitable for the secondary use. In this way the principle actually works in favour of data reuse. As explained above, data reusers are also considered data controllers and consequently they have to comply with all data protection requirements, including the data quality principle. The system of licences provided by Article 8 of PSI Directive is a good tool to reinforce data protection by further reusers as well as to help them clearly define responsibilities of the data controllers (LAPSI, 2012, p. 20).
- The security safeguards principle states that reasonable precautions should be taken against risks of loss, unauthorized access, destruction, etc. of personal data. Similarly to the OECD approach, General Directive states no concrete safeguards, but leaves this area up to MSs to regulate.[22] As the regulation is vague and scarce, data controllers, especially those who are more exposed to the media attention and more likely to be screened by the supervisors, are seeking for new forms of regulation. International industry standards, such as the ones provided by the International Organization for Standardization (ISO)[23], are getting more popular. It should be noted that in terms of data reuse the security of data is a considerable risk for data controllers: during the secondary use the data may be disclosed (or otherwise challenged) and, although the controllers are not directly responsible for the disclosure, they remain accountable to the data subjects. The security principle is strongly related to the accountability principle. The latter reinforces security requirements by encouraging data controller to implement practical tools for effective data protection (Article 29 Working Party, 2010). By shifting the responsibility to the controller, this General Directive's provision is aimed at achieving active compliance of controllers (European Union Agency for Fundamental Rights & Council of Europe, 2014, pp. 72–75).
- The openness principle states that the data subject should be able to know about the existence and nature of personal data, its purpose, and the identity of the data controller. According to the General Directive, data subject needs to be given the relevant information in an intelligible

[22] Now the EU is turning to a stricter and more harmonized regime of cyber- and data security. In 2013, as part of Digital agenda for Europe, the EC proposed a Directive of the European Parliament and of the Council concerning measures to ensure a high common level of network and information security across the Union (also known as NIS Directive) to raise the standards of data security in Europe. See more in section 3.2.4.8.

[23] The ISO 27000 family of standards focuses on the security of information assets. Using the standards helps the organization manage the security of assets such as financial information, intellectual property, employee details or information entrusted to you by third parties. ISO/IEC 27001 is the best-known standard in the family providing requirements for an information security management system (ISMS). More information on ISO standards available at: http://www.iso.org/iso/home.htm.

form including the details of: purposes of processing, the categories of data concerned, the data undergoing processing, the recipients or categories of recipients to whom the data are disclosed, any available information about the source and logic involved in any automatic processing of data (European Union Agency for Fundamental Rights & Council of Europe, 2014, p. 112). Data controllers and processors usually provide this information in their privacy policies or statements that are publicly available. However, the way in which this information is transmitted to the data subject is often too complex and lengthy to be fully and correctly comprehended. This is the reason that the operationalization of the openness principle has been subject of relentless criticism. For instance, Schermer, Custers, and van der Hof, Simone (2014) establish that privacy policies nowadays contain information overload, absent a meaningful choice for the users which leads to the situation where data subjects no more make informed decision but simply consent whenever they are asked to do so. Not only are data subjects unaware how their information and under what conditions will be processed, they also lack some basic understanding of whether their data can be and will be reused.

- When personal data is transferred to a third party and reused, the openness principle does not cease to apply (neither do other data protection principles). On the contrary, at this point it becomes even more important that the data subject is fully informed about the activities in which his or her data is involved. There are two options how to secure the data subject's consent and ensure his awareness. First, data reuse activities that might happen in the future are described and communicated to the data subject before his personal data is collected. Secondly, the data subject renews his consent every time before the data is reused in a new purpose. Both tactics prove to be difficult to apply. In the first case, it is hard to predict all the purposes for data reuse that may appear in the future. In the second case, it is almost impossible to get in touch with all data subjects and to secure their valid consent.
- In addition to the openness principle, the individual participation principle establishes the data subject's right to have his personal data erased, rectified, completed, or amended.[24] As a part of the right to obtain information or to demand erasure, General Directive obliges the controller to notify the third parties, to which the data was transferred, about the data subject's request, unless this would involve disproportionate efforts (European Union Agency for Fundamental Rights & Council of Europe, 2014, pp. 72–75). In addition, according to the General Directive, a data subject has the right to object processing of his data whenever they could be used in direct marketing purposes, be subject to automated decision (such as credit scoring) or when they compel with some specific interests of the data subject (European Union Agency for Fundamental Rights & Council of Europe, 2014, pp. 72–78). As data reusers are also considered data controllers, they should abide with the same data protection obligations, including the

[24] Note that in the General Directive, this principle applies only to incomplete or inaccurate data, or data that are irrelevant or processed illegitimately.

provisions on data subject rights – right to access to data, right to rectification, right to object to processing, right to erase/block the data (LAPSI, 2012).[25]

The underlining idea of the abovementioned principles and rules is to protect personal information, particularly those that are most sensitive such as health data, ethnical data, data indicating sexual orientation or religious views. While this approach clearly empowers the citizens with the sense of safety and trust, it means a barrier to data reuse. As LAPSI researchers (LAPSI, 2012) realized in the public sector, the full application of the personal data protection directive to the reuse of PSI could seriously limit the possibility of re-using the data that contains personal data and oppose the ideas expressed in the Recital 5 of the Directive on PSI that "[…] wider possibilities of re-using PSI should inter alia allow European companies to exploit its potential and contribute to economic growth and job creation." Similar issues have been identified in the private sector. Miani et al. (2014, p. 68) has pointed out that data protection regulation and associated issues such as privacy and ownership can critically restrict access to medical information for healthcare industry.[26]

3.2.4.1.4 Applicability of the directive

In terms of its personal scope, the General Directive (Article 3 (2)) excludes itself from the processing operations that fall outside the boundaries of EU law (e.g. operations concerning public security) or to processing of data by a natural person in the course of purely personal or household activities.

As regards its territorial scope, General Directive provides for two (main) legal bases for its application in EU MSs. Firstly, as Article 4(1)a foresees, EU law should apply whenever a data controller is established

[25] Reusing personal data accumulated from social media, for example personal twits, has become a popular business model. Small and medium-sized enterprises (SMEs) that do not own the data themselves are able to get the insights into the trends and analyze their (potential) costumers' preferences by collaborating with big social media providers. The activities are regulated by standard terms and conditions imposed by large data owners e.g. Twitter, Facebook. These conditions also reflect some of the privacy requirements. For example, there is a condition that requires third parties to cease using the twits that have been deleted by the users as soon as Twitter informs them about the deletion. It is unknown, however, what the level of compliance to this kind of requests is and how this requirement affects the business, especially SMEs and start-ups. Given the low effectiveness of DPAs in enforcing data protection rules, it is unlikely that data protection law would hinder reuse of data due to unwillingness of the reusers to accept their legal responsibilities related to personal data.

[26] Luckily, not only problems were identified, but a few solutions have been proposed, too. Public sector databases and registries could also include a kind of technical system that would help public bodies to anonymize personal data after the storing time of their first processing in order to automatically allow reuse of these data after this anonymisation. This solution should meet the national legislations that already impose total anonymisation of identities (e.g. Belgium), but the questions still remain whether it is technically feasible and would it allow a kind of *interoperability* between systems LAPSI (2012, p. 20). In the private sector, Miani et al. (2014, p. 63) suggests involvement of a trusted third party, depersonalisation tools and electronically run consent management.

in the EU and processes the data in the context of the activities of establishment.[27] Secondly, it should also apply when a non-EU data controller uses its equipment in a MS (without being established on the European territory). Apparently, the second rule tries to prevent the evasion of data controllers of their legal responsibilities through relocation of their establishments outside the EU, while still using technical means located in the EU to processes data in a way inadequate for the European standards (Kuner, 2007, p. 112).

3.2.4.1.5 International transfers

The regulation of international data transfers is one of the striking areas in data protection law, particularly for corporations operating internationally. In the era of globalization, data is being daily transferred all over the world, but it is evident that not all the countries provide an adequate level of data protection. Article 25(1) of the General Directive allows for the transfer of personal data outside the EU borders only if the third country ensures an adequate level of protection. It is up to the Commission's and MS' decision which countries they consider adequate. So far, the Commission has issued 13 adequacy decisions confirming adequate status of the following countries: Andorra, Argentina, Australia, Canada, Switzerland, Faeroe Islands, Guernsey, Israel, Isle of Man, Jersey, US[28], New Zealand ad Uruguay.[29] In addition to adequacy decisions, Article 26 of the General Directive provides two other legal bases for international data transfers. The first base are standard contractual clauses that exist in two different forms – as standardized set of clauses approved by the Commission or as ad hoc contracts that are individually approved by a national data protection authorities. The second base are exceptions such as consent of a data subject or public interest which occasionally legitimate data transfers. Despite not being explicitly mentioned in the General Directive, binding corporate rules (BCRs) are another useful tool to facilitate data flows (and ensure compliance). The Working Party 29, the European data protection body, has first recognized these corporate privacy policies that regulate data flows within one organization and tend to be *internally binding* within the organisation on all group companies and employees and *externally binding* for the benefit of individuals i.e. must create third party beneficiary rights for the individuals (2011, p. 186).

[27] In the attention-grabbing Google Spain case, the referring court asked the CJEU whether Google Spain advertising-related activities mean processing of personal data carried out in the context of the activities of an establishment of the controller on the territory of a Member State. The Court decided for a broad interpretation of the directive and stressed it would be contrary to the objective of the provision in Article 4 — which is to ensure, through a broad definition of the concept of 'controller', effective and complete protection of data subjects — to exclude the operator of a search engine from that definition on the ground that it does not exercise control over the personal data published on the web pages of third parties. Case C-131/12 from May 13, 2014, para. 21, 34 and 56. See also van Alsenoy and Koekkoek (2015).

[28] Through the Safe harbour framework.

[29] Information is available at the European Commission's official website: http://ec.europa.eu/justice/data-protection/document/international-transfers/adequacy/index_en.html.

Apparently, if the EU wants to establish a sustainable data reuse economy, reuse of data cannot be restricted to European countries. International transfers of data are preconditions for selling or acquiring the data to/from abroad. If this data also contains personal information, it is necessary to take additional precautions and use the tools such as standard clauses, BCRs or safe harbour.

3.2.4.1.6 Reform of data protection law

Finally, it should be noted that the EU data protection law has been the subject of ongoing negotiations. The legislative process started in 2012 when the European Commission (EC) presented its first proposal of the EU regulation on data protection.[30] The objective of the new law was to strengthen data protection and adapt it to the changed circumstances in the globalized and intra-connected world. The fact that the Commission proposed to replace the existing directive with a regulation means it was seeking for a complete harmonisation i.e. unification of all MSs' national legislations in this area.[31] The Commission emphasized that a unified approach would increase the standards of data protection and speed up the transfers of data across the EU (European Commission, 2012).[32] In March 2014, after intensive negotiations in the Parliament, the initial Commission's proposal was significantly amended.[33] Until June 2015 the negotiations were taking place in the Council. It is worth noting that there was no time limit for the Council's first reading, which explains why this took such a long time, especially since the EU MSs disagreed amongst themselves on some of the key changes, for instance on the so-called *one-stop-shop* rule (Proust, 2015). On June 15, the Council reached a general approach on the general data protection regulation.[34] In June the negotiations have started in a form of a trilogue between the Commission, Parliament and the Council. The aim of the trilogue process is to achieve agreement on a final text of the Regulation by the end of 2015 (Hunton & Williams, 2015).

Although three versions of the proposed regulation have been already shared with the public (the EC's original proposal, the Parliament's amendments and the Council's general approach), it is difficult to predict the final outcome and its consequences. However, many authors argue the proposed regulation represents a *Kopernican movement* in the EU law and predict that it will have a massive impact on European companies.[35]

---

[30] On February 25, 2012, the Commission publicly revealed the draft regulation and published additional materials to support the proposed reform including its communication and impact assessments: http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm.

[31] Regulation has a direct effect which means it has to be directly transposed into the national legal systems.

[32] The data protection regulation proposal was one of the first steps toward the EU single digital market. This initiative has been recently restated and reaffirmed as one of the main goals in the next Commission's mandate (more information on the EC website: http://ec.europa.eu/priorities/digital-single-market/).

[33] They added more than 4000 amendments Proust (2015).

[34] The most notable Council's positions relate to the fines (decreased to 2% of annual revenue), one-stop-shop provision and to the notion of establishment. More information is available at the following website: http://www.consilium.europa.eu/en/press/press-releases/2015/06/15-jha-data-protection/.

[35] See for example Kuner (2012, p. 4) and Lynch (2013).

For the purpose of this document, we have analyzed the proposed changes based on the Commission's proposal. Although this is not the most recent version of the regulation available, it is the only official and complete draft that have been publicly shared so far. The novelties from the proposed regulation, which are likely to influence data reuse practices, are listed and assessed below. It should be borne in mind, however, that despite the growing media attention that accompanies the legislative procedure, some renowned authors have stressed that the majority of the changes the proposed regulation makes are incremental at best (Rubinstein, 2013).

New definitions: The Commission's proposal gives new meanings or amends some of the key definitions from the currently used General Directive. For example, in Article 9(1) it expands the definition of sensitive data to include genetic data and data concerning "criminal convictions or related security measures". What is fundamental to note is that the concept of identification will likely no longer be limited to the possibility of knowing the address, name, etc. of an individual, but rather will focus on the likelihood of *singling out* an individual whether directly or indirectly (McKean, 2014). The broader definition of personal data means that the amount of information that can be reused limitless has been restricted and that reusers should take additional precautious when processing the data that might contain some personal information.

Explicit consent: The draft regulation suggests a tighter definition, namely an explicit consent, which will presumably leave less space for circumventions of the rules and put a bigger burden on the controllers (Kuner, 2012). Article 4(8) presupposes that such consent will be obtained either by a statement or by a clear affirmative action of a data subject. Given the fact that the consent is not a very feasible option to legitimate data reuse (see section 3.2.4.1.3), explicit consent requirement will close the door to this legal basis to an even greater degree.

Principle of transparency and accountability: By introducing the principles of transparency and accountability the EC confirms its efforts to raise the European standards of data protection for individuals. Article 22(3) stipulates that compliance measures should be independently verified, through the use of "independent internal or external auditors", however, this is only required when "proportionate". The concept of accountability seems to include the measures listed in Article 22(2) such as keeping documentation of data processing, implementing data security requirements, notifying data breaches, performing data protection impact assessments and designating a data protection officer (Kuner, 2012). Data reusers have the status of data controllers and consequently all these measures are also applicable to them. The difficulties that they might encounter when striving for compliance with the provision are described in section 3.2.4.1.3.

Right to be forgotten: the Commission's proposal reaffirms the data subject's right to have data deleted from the General Directive by stressing everyone should be able to request deletion of personal data if it is not necessary anymore in relation to the initial purpose (European Commission, 2014a). As Hoboken (2013) observes, the added value of the updated provision for data subjects that want to see

their data deleted is relatively minor, though it is the improvement that has received the greatest attention in academia as well as from general public. Kuner (2012) on the contrary believes the new provision is an important one, in particular its reversed side, the duty of the controller to inform third parties about the data subject's request to erase data. This duty, which is stipulated in Article 17 of the draft regulation, is limited to what is possible and does not involve a disproportionate effort. Similarly to other data subjects' rights, the right to be forgotten might turn out to be difficult to comply with. Some of the difficulties are mentioned in section 3.2.4.1.3.

Right to data portability: Data portability has been introduced into the proposal on data protection regulation as a new concept that aims to address the competition law concerns (for instance the lock-ins observed in market behaviour of cloud computing providers) in order to improve competition in the market, so that new services can innovate and attract customers away from the original device or service. Geradin (2014) defines it as the ability for people to reuse their data across devices and services and, therefore, it can be claimed data portability is in fact data reuse *per se.* The proposal for General Data Protection Regulation splits the right to data portability into two elements. Firstly, under Article 18(1) of the draft Regulation, individuals, whose personal data is processed electronically and in a "structured and commonly used format", are given the right to obtain a copy of that data for further use. Secondly, Article 18(2) provides for the right for individuals to transmit their personal data from one provider to another. Given that the majority of personal data are processed electronically, Article 18 seems to have widespread applicability and the potential to offer significant benefits to individuals (Bapat, 2013). Kuner (2012) sees the main advantage of this new right in enabling the individuals to change online services more easily by granting them the right to obtain a copy of their data from their (former) service provider. Zanfir (2012) joins his view, stating data portability as proposed by the EC presents a safeguard for data subjects' human rights. Swire and Lagos (2013) oppose. They warn that the concept of data portability in the proposal is over-broad, applying to small enterprises, to enterprises with no monopoly power, and to markets with no barriers to entry. In addition, they believe the EC went too far by proposing data portability a new human right. There has been no well-defined or established right to data portability and no jurisdiction has ever experimented with anything resembling it.

Higher penalties: Under the General Directive administrative sanctions were left to implementation by the MSs and consequently the numbers varied widely. The proposed regulation plans to transform the old system by substantially increasing the sanctions over what was previously possible and unifying their rate across the EU. The sanctions are imposed mandatorily for any intentional or negligent violation of certain provisions of the Proposed Regulation, and are divided into three categories, amounting up to

2% of a company's annual worldwide turnover (Kuner, 2012).[36] Increased administrative penalties impose additional risk on data reusers and force the data-driven companies to operate very carefully.

Lead authority: Article 51 of the proposal provides that a company (acting as a controller or processor) established in more than one MS, is required to approach the authority in the state of its main establishment, the so-called lead authority. Thus, instead of dealing with 28 different data protection authorities, the company has only one respondent, which importantly simplifies privacy procedures. If there is a lawsuit brought against a company, this should be done in the state of the lead authority. This provision strongly interferes with the MSs' jurisdictional regimes and has, hence, received least favourable feedback from the Council. Experts that work closely to the members of the Council have told this has been the point where the negotiations were blocked and intensive discussions were going on (Peers, 2015). The Council's finally reached the agreement by decreasing the powers of the lead authority to cases where the processing "substantially affects or is likely to affect substantially data subjects in more than one Member State".[37]

3.2.4.1.7 Data protection around the world

As explained in section 3.2.4.1.1, the OECD Convention 108 meant a turning point for data protection laws. Since 1980 a number of countries has started regulating the area and introduced its own laws. Europe has taken the leading role in this regard, but data protection has also been an important issue in many other countries.

In the US, data protection does not have such a prominent role in the constitutional acts as in the EU, where it is explicitly guaranteed in Article 7 of the EU Charter on human rights. The US data protection law is mostly focused on regulation of specific areas, such as health[38], financial services[39] and e-commerce[40]. Federal trade commission acts as a data protection body using its general authority to prevent unfair and deceptive trade practices (DLA Piper, 2015). There is no general data protection law on the federal level. The states can regulate certain issues on their own, for example whether it is necessary to notify data breaches or whether the companies need to appoint a data protection officer.

In Canada, the level of restrictiveness is similar to the one in the EU. There are 28 federal, provincial and territorial privacy statutes that govern the protection of personal information in the private, public and

[36] Up to 5% according to the Parliament version. See European Parliament legislative resolution of 12 March 2014 on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD)).
[37] The comparison between the Parliament's text and the Council's version is available via https://edri.org/files/EP_Council_Comparison.pdf.
[38] For example Health Insurance Portability and Accountability Act of 1996 (HIPPA).
[39] For example The Gramm–Leach–Bliley Act (GLBA), also known as the Financial Services Modernization Act of 1999.
[40] For example Controlling the Assault of Non-Solicited Pornography And Marketing Act of 2003 (CAN_SPAM Act).

health sectors. Although each statute varies in scope, substantive requirements, and remedies and enforcement provisions, they all set out a comprehensive regime for the collection, use and disclosure of personal information (DLA Piper, 2015).

Australia has also considerably robust data protection laws, while the regulation in the rest of the world is more lenient. In Africa, Asia and Latin America, there are a number of countries that have no laws on data protection (DLA Piper, 2015).

In recent years, we can observe that countries try to regulate data protection on a transnational level to simplify international data flows. The EU and the US use safe harbour as a mechanism to facilitate data flows between the continents. In 2011 Asian Pacific Economic Cooperation (APEC) introduced its cross-border privacy rules system, a principles-based tool to encourage the development of appropriate information privacy protections and ensure the free flow of information in the Asia Pacific region.[41] In 2014 African Union adopted its convention on cyber security and personal data protection, which means an important step toward better regulation of data protection in African regions.[42]

3.2.4.2 Privacy law

EU PRIVACY LAW	
Primary EU law	Charter of the fundamental rights of the EU (Arts. 7 and 8) Treaty on the functioning of the EU (Art 16)
Secondary EU law	General Data Protection Directive Regulation concerning the protection of individuals with regard to the processing of personal data by Community institutions and bodies Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (hereafter E-privacy directive) General Data Protection Regulation (draft proposal)

[41] The text is available via http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~/media/Files/Groups/ECSG/05_ecsg_privacyframewk.ashx.
[42] More information available on http://pages.au.int/infosoc/cybersecurity.

Personal data protection law is closely related to but not the same as privacy law. Privacy has several different aspects, such as spatial privacy (for instance, in your home), relational privacy (for instance, during phone calls), physical integrity (for instance, not to be touched without consent) and informational privacy (for instance, the use of personal data). Personal data protection law particularly sees to informational privacy. However, big data developments may also affect to some extent the other types of privacy.

Before discussing this, it is important to distinguish the concept of privacy and the right to privacy. The right to privacy cannot be claimed in all situations. There are limits to the right to privacy and in particular situations the right to privacy can be derogated. An example of this may be authorized police searches, such as frisking, home searches and wiretapping. From a legal perspective (and usually from the perspective of the individual involved as well) such actions violate someone's privacy and right to privacy. However, such violations are allowed under particular circumstances (such as reasonable suspicion and court orders) by priority rules for police forces.[43]

There does not exist a universally accepted definition of privacy. Warren and Brandeis (1890) defined the right to privacy as the right to be left alone. This definition is still used in the United States to some extent. Within the EU the right to privacy is explicated in article 8 of the European Convention on Human Rights (ECHR), which protects "private and family life" and "home and correspondence" and in articles 7 and 8 of the Charter of fundamental rights of the EU.

How do big data developments affect (the right to) privacy, other than personal data issues described in the previous section? As explained above, personal data protection mainly focuses on a set of principles for the fair processing of information. As such, personal data protection (1) focuses on the process (i.e., information processing rules) rather than on the consequences (i.e., potential negative aspects for privacy) and (2) focuses on personal data rather than other kinds of data that may (also) cause privacy issues.

In cases in which the legal regime for protection of personal data is considered too restrictive, data controllers increasingly seem to make use of aggregated, anonymised data. Such data can be as useful as personal data in many cases. A typical example may be a company that wants to personalize its marketing campaigns with the help of profiling. The use of personal data may be helpful to assess which people are potentially interested in particular products or services, but aggregated data on street level or neighbourhood level may be similarly useful and cheaper to process (no consent procedures required, no too detailed selection procedures necessary). For more examples, see Zarsky (2003). The fact that such targeting is not completely accurate (false positives and false negatives may exist) does not matter significantly and outweighs the costs for a more accurate approach.

[43] In legal terms such rules are called *lex specialis* (special rules) that have priority over *lex generalis* (general rules).

Anonymised data are not protected by the General Directive. However, it is increasingly easy to de-identify anonymised data (Ohm, 2010). Furthermore, the use of big data further enables (de-)identification of data. Such predictions can be made with high accuracy: Kosinski, Stillwell, and Graepel (2013) showed how a range of highly sensitive personal characteristics, including sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances and parental separation can be predicted very accurately on the basis of Facebook likes. Obviously, predicting missing values is also possible for people who (whether on purpose or not) provided false information.

These practices of profiling, personalization and (de-)identification may have negative consequences for privacy. For instance, the meaning of informed consent changes. Characteristics of people who refused to provide consent to process their personal data may be predicted anyway with the help of big data. Furthermore, the number of consent requests that people are confronted with nowadays is so large that most people do not have time to really read (let alone properly understand) all the terms and conditions each time they provide consent (Schermer et al., 2014).

Another aspect of big data is that it is more difficult to provide (intuitive) transparency on the ways in which data is processed. Several tools for data analyses, notably data mining tools are complex and may be difficult to explain to data subjects. Furthermore, data mining tools are designed to extract previously unknown knowledge from datasets, which means that their outcome is by nature unpredictable to some extent. These aspects make it very difficult to provide transparency about the ways in which data is processed. A lack of transparency about data processing may result in outcomes and decisions about people that may be unexpected or unexplainable. Solove (2006) states that such privacy issues are more a Kafka problem (a lack of transparent decision making) than a Big Brother problem (constant monitoring of people).

Furthermore, a major issue regarding profiling is the fact that the patterns resulting from such data analyses may turn out to be stigmatizing or discriminating. This will be discussed in the next section.

To wrap up, the use of big data may have negative effects on (the right to) privacy and (as we will explain in the next two sections), equal treatment/non-discrimination and other human rights such as human dignity. Data reuse may increase these issues and therefore there may be reasons (on either moral or legal grounds) to limit data reuse of some types of data, for some types of processing and for some purposes of data use and reuse.

3.2.4.3 Non-discrimination law

EU ANTI-DISCRIMINATION LAW	
Primary EU law	European Convention on Human Rights (Art. 14) Charter of the fundamental rights of the EU (Arts. 7 and 8) Treaty on the functioning of the EU (Art 16) United Nations' documents related to human rights
Secondary EU law	Anti-discrimination directives

Big data also raises issues with regard to equal treatment and non-discrimination and related laws. As indicated above, big data may be useful for profiling purposes, but the results from profiling and other types of data analyses may turn out to be stigmatizing or discriminating. When selecting individuals or groups of people on particular characteristics, this may be unwanted or unjustified or both. Selecting for jobs, offering products and services to specific groups only, and some other decision-making is considered unethical and, in many countries, forbidden by (anti-discrimination) law when it takes place on the basis of gender, ethnic background, etc. When risk profiles constructed by companies, governments or researchers become 'public knowledge', this may also lead to stigmatization of particular groups. Discrimination and stigmatization on a large scale may also result in polarization of (different groups of) society.

Interestingly, research has shown that removing sensitive attributes (such as ethnicity, gender, etc.) from databases in order to prevent unethical or illegal discriminating results does not prevent finding such profiles (Kamiran & Calders, 2009). There are several possible explanations for this. For instance, most data mining tools make predictions on the basis of training data. If the training data is biased towards particular groups or classes of objects, e.g., there is racial discrimination towards black people, the learned model will also show discriminatory behaviour towards that particular community. These are self-fulfilling prophecies. Another reason is that often other attributes than the sensitive attributes that were removed will still allow for the identification of the discriminated community (Pedreshi, Ruggieri, & Turini, 2008). For example, the ethnicity of a person might be strongly linked with the postal code of his residential area, leading to a classifier with indirect racial discriminatory behaviour based on postal code. The postal code then serves as a proxy for ethnicity. This is closely related to the ease of making predictions of missing attributes in big data settings: this can be done for both identifying data items (resulting in privacy issues) and for items like ethnicity, religion, gender, etc. (resulting in discrimination issues).

3.2.4.4 Other human rights

Apart from privacy and discrimination there are other human rights that may be affected by big data developments and data reuse. Examples are human dignity, liberty and justice. Related moral values that may be affected are solidarity, autonomy and individuality. It is beyond the scope of this deliverable to describe in depth the implications of data reuse in big data and open data settings for these values. We will limit this section to human dignity and refer to existing literature for more on the other human rights and moral values.[44]

Human dignity is usually not considered so much as a human right but rather as the core of human rights. The Charter of fundamental rights of the EU starts with human dignity in its first article, stating that human dignity is inviolable and must be respected and protected.

Big data can put pressure on human dignity. Solove (2006; 2007) argues that in our information society, the reputation of people is more and more constituted by the data that is disclosed about them. Such disclosure of personal data can be voluntary or involuntary. As a result, people are also increasingly judged upon their digital representation (the digital person) rather than human beings of flesh and blood. This may be particularly problematic when characteristics of digital identities are incorrect or incomplete. It may also be problematic when automated decisions (i.e., without further human interference) are made upon individuals based solely on their digital identity.[45] Practices like profiling described above can reinforce a tendency to regard persons as mere objects (Bygrave, 2002). Another issue may be so-called chilling effects. This refers to the fact that people may alter their behaviour when they are aware that they are being monitored. Sometimes, for instance in cases of camera surveillance, the aim is precisely to make people behave 'better', but a more general effect may be that people behave more modest and reluctant overall, reducing their freedom of expression, liberty and other important human rights and values.

3.2.4.5 Data retention laws

EU DATA RETENTION LAW	
Primary EU law	Treaty on the functioning of the EU (Art 105)
Secondary EU law	Data Retention Regulation (proclaimed invalid)

[44] See, for instance Vedder (1999). Autonomy and liberty in the context of Big Data are often related to consent (see previous sections). Justice, solidarity and individuality are often closely related to the discrimination issues mentioned in the previous section.

[45] Note that EU personal data protection law prohibits automated decision-making that is solely based on automated processing of data. See art. 15 of General Directive.

In the post-9/11 era the governments have exhibited greater ability and desire to develop a mass surveillance system, whereby a common trend has been intensive monitoring of citizens' communications and online behaviour (Vedaschi & Lubello, 2015). Fight against serious crime, especially terrorism, justified the intensified collection of communication data and extended data retention periods all over the world. These new legal measures, in spite of seeking legitimate objectives, have raised some serious concerns, mostly related to the right to data protection and privacy. In the EU the Directive 2006/26 aimed to harmonize the MSs' laws to ensure a minimum period of retention as well as the minimum scope of retained data. The directive imposed an obligation upon internet service providers (usually private companies) to collect and store, for a significant period, a large and varied amount of metadata, which may be of use to public security authorities in the fight against serious crime (Vedaschi & Lubello, 2015). National implementation laws introduced different solutions to fulfil the directive's requirements, but a number of them were later brought to the national constitutional courts, due to alleged violations of the fundamental right to privacy and data protection (see for example Rauhofer and Mac Sithigh 2014). In 2013 an Irish court referred to the Court of Justice of the EU (CJEU) with a request for a preliminary ruling on validity of EU data retention directive. The CJEU held that since the directive covers, in a generalized manner, all individuals, all means of electronic communication and all traffic data[46] without any relevance to the level of the crime (the directive was only adopted with the goal to fight serious crime and terrorism) and enables easy access to all sensitive information, it is a disproportionate measure and invalid in the light of Article 7 and 8 of the EU Charter on human rights.[47] The fact that the data retention directive was found invalid does not directly imply the implementation laws are automatically invalid. This needs to be assessed in every specific case, based on specific national law's provisions. Now the EU MSs have two options how to deal with the issue: either they repeal their entire national data retention legislation or they check whether it is compliant to the charter and amend it accordingly[48].

However, the debate on retention has not been closed yet. Soon after the retention directive was found invalid, the EU parliament started discussing the adoption of new anti-terrorism measures such as EU passengers name records to respond to the increased threat of terror, coming from the EU as well as from the outside (viEUws, 2015). Following the attacks on Charlie Hebdo journalists in Paris the German government proposed a new law on data retention which the opposition claims is only a relabelled

[46] "Traffic data means any data processed for the purpose of the conveyance of a communication on an electronic communications network or for the billing thereof." (Article 2 (b), Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector)

[47] Joint cases C-293/12 Digital Rights and C-594/12 Seitlinger from April 8, 2014.

[48] Committee on Civil Liberties, Justice and Home: Report on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD)). http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FTEXT%2BREPORT%2BA7-2013-0402%2B0%2BDOC%2BXML%2BV0%2F%2FEN&language=EN

former law that was earlier found disproportionate by their constitutional court (Tost, 2015). The final outcome of these legislative procedures remains to be seen, but they surely indicate the discussion on data retention has yet not come to the end.[49]

Data retention laws present an example of a regulation which rather encourages than restricts data reuse. Namely, its main purpose is to ensure there is sufficient amount of telecom data available to law enforcement for later reuse (under strict conditions). While fighting the crime is no doubt a valid argument for vast data collection and reuse, the example shows how human rights can be at stake. Moreover, the retention requirement imposes some significant administrative and financial burden to the communication providers, making it undesirable from the economic perspective.

Data retention laws described above should not be confused with data retention requirements that are part of industry-specific laws, codes of conduct or some other agreements. For example, Article 7 of the Code of Practice on Secondary Use of Medical Data in Scientific Research Projects (2014) provides that medical data is retained for a period of time that is needed to ensure reproducibility and verifiability of the findings. For pseudonymised and anonymised data the conditions are more lenient. Similarly, search engines' providers stick to the retention period as determined in their industry's best practices. In the EU, Article 20 Working Party has advised to limit the retention of personal data (search terms) to 6 months (2008, p. 19).[50] The retention period in private sector, which does law not require, has to comply with the principles of data protection law, in particular with the principle on data quality and on lawful processing.[51]

[49] Also in the US, the retention laws have been placed under judicial scrutiny. The Court of Appeal has recently found the bulk collection of communication metadata, cannot be justified under Article 215 of the Patriot Act (ACLU v. Clapper, 2015). Meanwhile, Obama administration's proposal of the Freedom act that would limit the collection and access to metadata is being discussed in the Congress.

[50] Interestingly, the economists argue Chiou and Tucker (2014) that longer data retention of historical data (i.e. users' search queries) does not represent an important competitive advantage, if measured by traffic on the search engines. So far, we have not come across a study that would directly measure data retention as a competitive advantage from the firm's (and not costumers') perspective measured by possible data reuse. It has been pointed out many times, however, that further processing of user (search queries) data touches on a core field of innovation of search engine technology and can have a high relevance for competition Article 29 Working Party (2013, p. 19).

[51] See more in section 3.2.4.1.3.

3.2.4.6 Data localisation laws

EU DATA LOCALISATION LAW	
Primary EU law	Charter of the fundamental rights of the EU (Arts. 7 and 8) Treaty on the functioning of the EU (Art 16)
Secondary EU law	Directive 95/46/95 (Chapter IV) Regulation (EC) No 45/2001 Directive 2002/58/EC General Data Protection Regulation

In the post-Snowden era, the idea of a transnational Internet where the information is freely moving cross-border has been recently challenged by a number of legislative proposals containing provisions on data localization. Data localization requirements refer to laws (or, parts of laws) that limit the storage, movement and/or processing of data to specific geographies and jurisdictions or that limit the companies that can manage data based upon the company's nation of incorporation or principal sites of operations and management (Hill, 2014). A striking number of countries has been moving closer to this paternalistic approach to cross-border data transfers by adopting various legislative measures with a common characteristic to encumber the cross-border data transfers (Chander & Lê, 2014). These measures are usually not implemented as a completely new set of rules in a country, but simply as provisions in (amendments to) the existing laws, requiring, for instance, the data to be only located on home country servers. Also, data localization may be reflected in a less direct form such as a requirement forcing local hiring or local purchasing of the information and communication technology (ICT) equipment (Hill, 2014).

Governments have different motives to take a stricter approach to data transfers. However, there are a few points common to all these initiatives. At the list of possible motives, Choi, Dooley, and Rungtusanatham (2001) place first fear of foreign surveillance, followed closely by privacy and security concerns, endeavours for better law enforcement and development of national economies. Other possible motives, although they are harder to justify, are protectionism, information control, limitation of free speech and anti-globalist populism.

While the listed reasons could be understandable to a certain degree in the light of Snowden's disclosure, they have been widely criticized from both the legal and economic perspective.[52] According to Chander and Lê (2014, pp. 28–29) , foreign surveillance cannot be stopped by localizing data, on the contrary, it can actually be made easier due to less rigorous locally–based security systems. Similarly,

[52] See for example Chander and Lê (2014), Bauer, Lee-Makiyama, Marel, and Verschelde (2014), Hill (2014).

localisation requirement would hinder economic growth, since this would mean local firms, especially start-ups, would have no access to cheap and high-tech data services, for instance cloud computing or web mails. The economic impact has been estimated to be considerably high: according to ECIPE researchers (2014), data localisation could decrease a national GDP up to 1.7%.

Chander and Lê (2014) identified 13 countries where data localisation laws (or at least data localisation provisions) have been either enacted or proposed. In Australia it is prohibited to transfer electronic health records abroad, unless one of the few exceptions applies. In Brazil the newly proposed bill on digital rights and internet (Marco Civil da Internet) includes an ability of the government to force Internet providers to store, manage and disseminate the data within the country. In Canada, a number of regional laws require personal information held by public institutions to be stored and accessed exclusively in Canada. In China, the government guidelines prohibit the transfer of personal data abroad without an express consent of the data subject or explicit regulatory approval. The EU limits data transfers to the so-called adequate countries or allows them only if sufficient safeguards are in place. Among its MSs, France and Germany are the countries where the idea of localization has been promoted most vocally. Ms Merkel, German chancellor, has encouraged the idea of building out a European Internet infrastructure and in 2013 Deutsche Telecom launched *E-mail made in Germany* that would keep German e-mail communication within national borders (2014, p. 11). In France, the idea of promoting local data infrastructure was developed even further by proposing a new tax on collection, management and commercial exploitation of personal data generated by users located within the country (Pfanner, 2013). However, this idea has never become law. Yet France has taken some other concrete steps to protect the data of its citizens. Soon after the Snowden disclosure, it passed in law two bills: the first one restricting the establishment and implementation of the systems for interception of electronic communications and the second one, the Military Programming Law, giving permission to the national security and intelligence services to see electronic and digital communications in *real time* Chander and Lê (2014). Another significant example of data localisation has been Russia with its amended draft law On Personal Data and On Information, Information Technologies and Protection of Information stating all personal data of Russians to be collected and recorded on the servers in Russia (Shashina, 2014). A number of other countries have also proposed data localisation laws: Vietnam, South Korea, Kazakhstan and Taiwan (Chander & Lê, 2014).

Paraphrasing Chander and Lê (2014, p. 1), we could say that by creating *Schengen zones for data*, data localisation laws are undermining the possibility of global services as well as the major new advances in IT. For instance, Chander and Lê (2014, p. 40) warn about the threats these regulations impose on cloud computing, Internet of Things and big data industry. At this point it should be added that data localization also means an important restriction to data reuse economy, where cross-border flow of information and unlimited access to data are one of the main enablers.

3.2.4.7 Freedom of information and open data – regulation of PSI

EU DATA LOCALISATION LAW	
Primary EU law	European Convention on Human Rights (Art. 10) Charter of the fundamental rights of the EU (Art. 42) Treaty on the functioning of the EU (Art 15 (1))
Secondary EU law	REGULATION (EC) No 1049/2001 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 30 May 2001 regarding public access to European Parliament, Council and Commission documents Directive on PSI Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information

Open data, as defined for the purposes of the EuDEco project, refers to open governmental records, which may be reused to create value for the citizens as well as for businesses. Open data is often replaced by or confused with one of the following terms: open government, freedom to information and right to access PSI. However, there is an important difference between these notions. We therefore begin this section with a short explanation of each of them. Next, we place them within the legal framework and show how they relate to data reuse.

The idea of freedom of information, also referred to as *freedom to access*, was introduced in the US after the World War II in reaction to governmental opacity and claims for more transparency. In 1966, when the Congress adopted the Freedom of Information act, open government (also open data), was used as a synonym for public access to previously undisclosed government information (Yu & Robinson, 2012). As such, it meant a measure to provide citizens with better control over the government and to ensure its accountability.

Throughout the 1970s several European countries (Denmark, Norway, France and the Netherlands) adopted Freedom of information acts. In the 1980s Australia, New Zealand and Canada Access joined and the trend has continued until today (Glover, Holsen, MacDonald, Rahman, & Simpson, 2006).

The advent of Internet put the idea of open government (open data) into a new light. As soon as the Internet opened to individual users in the early 1990s, government data started going online (Yu & Robinson, 2012, p. 190). Suddenly, *open government* became a label not only for political

accountability but also for technological innovation (Yu & Robinson, 2012, p. 193). Most notably, in 2008, immediately upon starting his first mandate, US President Obama took some significant steps toward an open and technologically progressive government by adopting the Open Government Directive.[53] Not only did the directive aim at higher government transparency and accountability, it also acknowledged that open data might create economic opportunities. However, critics claimed the new law only facilitated repackaging of old information providing *marginal value* to the economy. They urged the government to also make available those public data that hold the state accountable for its policy and spending decisions (Yu & Robinson, 2012, p. 198).

At almost the same time similar initiatives started taking place in the EU. In 2003 the Council of EU adopted the Directive 2003/98/EC on the reuse of PSI (hereafter PSI Directive)[54], which intended to harmonise the MSs' legislation in order to open the data to public access and encourage its reuse. The directive encourages the MSs to make as much information available for reuse as possible. It addresses material held by public sector bodies in the MSs, at national, regional and local levels, such as ministries, state agencies, municipalities, as well as organisations funded for the most part by or under the control of public authorities e.g. meteorological institutes[55]. Outside the EU scope, the United Kingdom (UK) was running a campaign on its own, setting up a public portal with open governmental data – data.gov.uk. A number of other MSs followed the British example, either on their own initiative or sponsored by the EU. In 2013, the PSI Directive was revised and its scope was extended to cultural institutions such as libraries (including university libraries), museums and archives (PSI directive, Article 3).

Before the directive was amended in 2013, the EC launched LAPSI, an international research initiative, which brought together EU legal experts to identify the remaining legal obstacles to access and reuse of PSI on the European content market, and to propose measures and tools to stimulate the progress of the European market towards open data.[56] The project identified a number of important legal barriers that impact the PSI reuse as well as made some welcome recommendations. The highlighted topics were: the protection of IPRs when embedded in the PSI, competition law restrictions in case of owning

[53] Text of the directive is available at: https://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf.
[54] [2003] OJ L 345/90
[55] http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information
[56] The project's website can be accessed via: http://www.lapsi-project.eu/.

the data that may be used to compete on the market, and personal data protection.[57] Furthermore, the researchers pointed out that non-harmonized national laws are also blocking data reuse.[58] Some of the findings of the research project have been incorporated in the amended directive e.g. the cap of cost charged for the access and reuse of information (PSI Directive, Article 1 (6)). Now the maximum is set on the marginal cost that incur due to a specific request (cost of reproduction and dissemination but no more cost of collection and return on investment).[59]

3.2.4.8 Cybersecurity law

EU CYBERSECURITY LAW	
Primary EU law	CFR Arts. 7 and 8 TFEU 16
Secondary EU law	Network and Information Security Directive (EC proposal)

As Hanover Research (2015) established based on their US experience, cybersecurity law has been a growing field of legal practice and a priority for legal practitioners. In the EU, the Commission set the basis of cybersecurity regulation with two communications: first, in 2011 with the Communication on critical information infrastructure protection and secondly, in 2013 with the Communication on Cybersecurity strategy for the EU. The Commission plans to realize this strategy with a Network and Information Security (NIS) Directive, proposed in early 2013.

The proposed directive aims to improve the security of the Internet and the private networks and information systems underpinning the functioning of our societies and economies. This will be achieved by requiring the MSs to increase their preparedness and improve their cooperation with each other, and

[57] With regard to charging for PSI, researchers think PSI should not be offered at a too low price, as this may harm the businesses opportunities of private undertakings that compete with public sector bodies in the generation of substitutable information (see LAPSI position paper no.1: The principles governing charging for reuse of public sector information). Also see the contributions by technologist and computer science experts, such as Robinson, Yu, Zeller, and Felten (2009). In particular, they refuse the idea of setting up dedicated websites for the public, or providing formats as under Art. 5 of the Directive, that entail an extra cost which can be incorporated in reproduction and dissemination charges should be double checked – and probably resisted – on the basis of the argument that public sector bodies should only provide *raw data* now, i.e. reusable data, not fancy web sites; and that in accordance with the engineering principle of separating data from interaction, public sector bodies should avoid seeking the best tools and leave to the market – and to market-based technologies – the optimization of the presentation of the data.

[58] For instance, in Slovenia, the PSI directive was implemented in 2003 and since then the public bodies have given open access to their administrative decisions, however, due to improperly written national guidelines or the absence of the European ones, the decisions can only be available in PDF format, which makes them rewritable and hard to be reused Salamanca and Eechoud (2014).

[59] For a detailed analysis see LAPSI position paper N. 1 Principles governing charging for reuse of PSI.

by requiring operators of critical infrastructures, such as energy, transport, and key providers of information society services (e-commerce platforms, social networks, etc), as well as public administrations to adopt appropriate steps to manage security risks and report serious incidents to the national competent authorities (European Commission, 2013a).

The Commission acknowledges that the Data Protection Directive already contains the rules on security standards for controllers (including reusers) of personal data. However, there have been no rules relating to those that control (or reuse) non-personal data. For example, a network and information security breach affecting the provision of a service without compromising personal data (e.g. an ICT outage at a power company resulting in a blackout) would not have to be notified (European Commission, 2013a). The NIS directive fills that gap by requiring market operators to notify critical breaches to national authorities and to ensure an adequate level of security for their information assets (Article 14, para. 1 and 2).[60]

3.2.4.9 National security law

Not harmonized area

National security is strongly associated with national sovereignty, where countries are free to impose their own laws and the EU legislative powers are limited. Nevertheless, most states restrict data reuse when it could jeopardize national interests and security. For example, research data pertaining to intelligence and military activities or the data related to nuclear, biological threats or human health systems should not be shared or reused (Mathae & Uhlir, 2012).

3.2.5 Private laws

Section 3.2.5 deals with private law rules that regulate data reuse. Every subsection begins with a table of applicable EU laws and continues with a general introduction to the chosen legal area. Subsections end with an analysis of the applicable rules through the prism of data reuse.

[60] The broad definition of a market operator in Article 3 of the NIS Directive (proposal) means that many data reusers would fall within the scope of the provision and would have to comply with the security standards as well as respect the notification duty.

3.2.5.1 IPR (rights in rem)

EU IP LAW	
Primary EU law	Treaty on the functioning of the EU, international treaties (e.g. The Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS))
Secondary EU law	Directive 98/84/EC on the legal protection of databases; Directive 2006/116/EC on the term of protection of copyright and certain related rights, replacing Directive 93/98/EEC; Directive 00/9/EC on the harmonization of certain aspects of copyright and related right in the information society ETC.

IPRs protect immaterial goods, which are mostly the product of a creative mental human activity in the industrial, scientific, literary and artistic fields (Kur & Dreier, 2013, p. 2). IPRs rights are recognized as *jura in rem,* rights "available against the world at large" (Black, 1910). The law grants to the holders of IPRs the legal power to use and to exclude others from using the immaterial good in question in any way as defined by the IPR legislation (Kur & Dreier, 2013, p. 2).

Based on the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) from 1994, which is the fundamental, internationally accepted set of IP related principles, we can distinguish seven groups of IP rights (Cook, 2010, p. 2):

- copyrights (and related rights) which protect literary, artistic and scientific works;
- patents granted for inventions;
- trademarks used to identify the commercial source of goods or services[61];
- industrial designs protecting the eye appeal of products;
- protection against unfair competition (also known as trade secrets);
- topographies;
- geographical indications.

Traditionally, copyrights, patents and trade secrets were considered models to which aspirants would turn for the protection of other ideas, information and trade values. When claimants were unable to fit a new subject matter within one of the model systems, they sought to have a new regime created to protect them. In this way, additional groups of IPRs emerged, such as industrial designs or trade secrets

[61] The directive on copyrights is currently in the legislative process and it will be presumably adopted in 2016. More information about the consultations and the legislative debate is available and regularly updated at the EC's official website: http://ec.europa.eu/internal_market/consultations/2013/copyright-rules/index_en.htm.

(2013, p. 10) that are nowadays a standard part of the IPRs family. This development has not finished yet. For example, some states have been considering a system of short-term rights protecting minor technical advances, e.g. the so-called petty patents. Also, there has been a pressure for expansion of copyright, creation of neighbouring rights and the protection of entertaining ideas (Cornish et al., 2013, p. 11). The disharmonized approach is typical for the EU, where several types of IPRs were introduced that are not the subject of TRIPS such as moral rights and the *sui generis* (database) right.[62]

Before data reuse takes place, it is necessary to consider whether this could violate IPRs contained within a dataset. For example, if a social network provider reuses the photos that have been made publicly available by its users, this can mean a copyright infringement, even though the photos are not creative. In Italy, Spain, Austria and Germany a neighbouring right for a non-creative work lasts between 20 and 25 years (Boom & Canova, 2013). [63]

Among all the IPRs, copyrights, database rights and trade secrets are most closely related to data. Patents can apply to software and business processes that manipulate and process data, but generally not in relation to data itself. Trademarks can apply to data products (like indices), but again, generally not in relation to the actual data (Kemp, 2014). In line with this view our analysis will focus on copyrights, trade secrets and the *sui generis* database right as it is evident these are the legal concepts that might block or at least defer data reuse activities in the EU.

3.2.5.1.1 Copyrights

Copyright protects the form or expression of information but not the underlying information itself. It applies to software, certain databases, literary works, music, films, videos and broadcasts.[64] Initially the

[62] Database protection in a form of a specific right only exists in the EU. This right is unknown, for example, in the US Derclaye (2008, pp. 2, 223).

[63] Normally, the social networks would receive a user's approval to reuse his or her (IP protected) posts as part of the pre-registration consent to the standard terms of use. However, the researchers from KU Leuven have established that, at least according to Belgian legislation, such a blank approval does not suffice as a justification for IPRs reuse. See more in van Alsenoy and Koekkoek (2015).

[64] Berne convention for the protection of literary and artistic works from 1886 incorporates a non-exhaustive list of copyright protected works: books, pamphlets and other writings; lectures, addresses, sermons; dramatic or dramatico-musical works; choreographic works and entertainments in dumb show; musical compositions with or without words; cinematographic works to which are assimilated works expressed by a process analogous to cinematography; works of drawing, painting, architecture, sculpture, engraving and lithography; photographic works, to which are assimilated works expressed by a process analogous to photography; works of applied art; illustrations, maps, plans, sketches and three dimensional works relative to geography, topography, architecture or science; translations, adaptations, arrangements of music and other alterations of a literary or artistic work, which are to be protected as original works without prejudice to the copyright in the original work; collections of literary or artistic works such as encyclopaedias and anthologies which, by reason of the selection and arrangement of their contents, constitute intellectual creations, are to be protected as such, without prejudice to the copyright in each of the works forming part of such collections. The complete text is available on: http://www.wipo.int/treaties/en/ip/berne/.

rationale for copyright protection consisted of protecting the investment made by publishers in the printing of books. Later, the focus shifted to the author (Kur & Dreier, 2013, p. 241). It has to be noted that a work only attracts copyright protection if it owns a certain degree of *creativity* (also *originality*) which EU law defines as "the author's own intellectual creation" (Kur & Dreier, 2013, p. 252). Copyrights arise automatically by operation of law in the EU (so no registration is required) and constitute a formal remedy that stops unauthorized copying (Kemp, 2014, p. 12). Copyright protection does not continue indefinitely. The law provides for a limited period of time during which the rights of the copyright owner exist. In the EU this period extends to 70 years after the author's death (World Intellectual Property Organization, 2008, p. 13).

The essence of copyright law is the idea that the author (or the owner of the copyright) has the power to authorize or to prevent certain acts in relation to a work (World Intellectual Property Organization, 2008, p. 7):

- reproduction in various forms, such as printed publications;
- sound recordings;
- distribution of copies;
- public performance;
- broadcasting or other communication to the public;
- translation into other languages;
- adaptation, such as a novel into a screenplay.

However, in specific cases the law allows exploiting the work without authorization. Free use is allowed in case of quoting from a protected work, or when a work is used to illustrate a case for teaching purposes. Another example are non-voluntary licenses, which allow exploiting the work without an authorization but do require that compensation is paid to the rights owner (World Intellectual Property Organization, 2008, p. 11).

In addition, the law grants to authors the so-called moral rights (opposite to the economic rights, which are the ones listed above): the right to claim authorship of the work and the right to object to any distortion or modification of the work. These rights remain with the author even after he has transferred his economic rights (World Intellectual Property Organization, 2008, p. 10).

Authors may transfer the rights to their works (apart from the moral rights) to individuals or companies best able to market the works, in return for payment. These payments are often made dependent on the actual use of the work, and are then referred to as royalties. Transfers of copyright may take one of

two forms: a non-revocable assignment and a revocable licence (World Intellectual Property Organization, 2008, p. 13).[65]

Copyright is an important concern for data reusers. In the context of data, the traditional literary copyright will mostly subsist in documentation – for example, the technical and user documentation relating to computer software and information architecture (Kemp, 2013). Any data analytics or data mining will often involve the wholesale copying of information or databases, all of which will be protected by IPRs in relevant jurisdictions (Graham & Lewington, 2013). Where data is not owned or licensed by the reusers, they will either need to abstain from using it or rely on an exception to copyright protection, otherwise they might be accused of violating a copyright. The reuse would constitute a copyright infringement under the following conditions (Kemp, 2014):

- that copyright subsists in the work – generally, that it is original and sufficient to warrant copyright protection;
- that the claimant owned that copyright;
- that the work was within copyright (life plus seventy years in the case of software, databases and other literary works); and
- that the copyright had been infringed – for example, a qualitatively substantial part of the work had been reproduced or reused without authorization.

However, it will not always be easy to determine who actually owns the data. The unambiguous ownership has been identified as a top-problem of open data use. In the case of CERN, the particle physics research institute, run by multiple international partners, they were unable to resolve which partner in the consortium owns which data. This caused a mismatch in their expectations around IP generated by the scientific research (2014, p. 12). Also in the private sector practice, Graham and Lewington (2013) have observed a gathering storm between data owners on the one side and technology providers on the other, as complex arguments relating to ownership, licensing and exceptions to copyright are being rehearsed. Kemp (2014) believes those problems are related to the uncertain scope of IP rights and the fact that the law in this area will surely continue to develop in the coming years as big data gathers pace (Kemp, 2014).

Concerns related to the manifestation of the copyright in the modern society have accounted to some vocal observations suggesting the copyright laws should be transformed to better fit the requirements of the digital economy. In the UK, the report on copyright reform prepared by Ian Hargreaves made sweeping and controversial recommendations for changes to copyright law. A key proposal was to permit non-commercial use of analytics as well as promoting at a EU level an exception to support text mining and data analytics for commercial use. This proposal was met with widespread criticism, particularly from copyright owners (Graham & Lewington, 2013).

[65] See more details on licences in section 3.2.5.2.

In 2015, mainly through the words of the EU Commissioner for Digital Economy and Society, Günther Oettinger, the road for a reform on copyright in the EU has been laid down. The proposed measures include the portability of legally acquired content, cross-border access to online services and greater legal certainty for the cross-border use of content for specific purposes such as research (European Commission, 2015, pp. 6–8). The EU Parliament has contributed to the copyright reform by providing a report on the implementation of EU Directive on the harmonisation of certain aspects of copyright and related rights in the information society. According to MEP's Sabine Verheyen, a member of European People's Party, the current debates have the potential to alter the European cultural sector in a lasting way. In order to thoroughly discuss the reform of the EU legal framework on sustainable copyright legislation, the EU Parliament has also set up working groups on the same topic (Verheyen, 2015).

In the academic sphere, some authors have been proposing to reform the system of IP rights even more profoundly.[66] They believe that the current setting is obsolete and that copyrights as we know them today (more precisely, the right of reproduction which is inherent to a copyright) should be transformed into a right to reuse. This would prevent the discrepancy between the social and legal norms that is most clearly seen on the Internet.[67] Namely, every online use of a work necessarily entails a reproduction thereof, so the reproduction no longer is regarded as a good indicator for infringement (Filippi & Gracz, 2013). This lucid analysis indicates that our society is willing to move toward a more open data economy where data reuse would outplace data reproduction, allowing for more aggressive but also more economically attractive intervention in IPRs.

3.2.5.1.2 Database right

Another concept that can play an important role in data reuse is the legal protection of databases. Countries have addressed the issue in an uncoordinated fashion using diverse legal mechanisms ranging from unfair competition rules to technological protection measures (Derclaye, 2008, p. 3). The European Community is an exception as it recognizes a unique right exclusively aimed at the protection of databases, the so-called *sui generis* right.

Although this right is obviously of its own kind, it has been most often classified as a right neighbouring copyright (Derclaye, 2008, p. 54). The policy idea behind database protection in the EU was mostly based on the idea of giving more economic incentives to database creators.

The new right was introduced in 1996 by Directive 96/9/EC of the European Parliament and the Council on the legal protection of Databases (hereafter Database directive). The aim of the directive as expressed by the EC was to remove existing differences in the legal protection of databases by

[66] See for example De Filippi & Grazc (2013), Ginsburg (2003).
[67] Legal norms advocate for restricted use and reuse of information, while social norms advocate for the free circulation of knowledge on the Internet. Thus, many activities (such as the practices of file-sharing, remix or mash-ups) are not perceived negatively by end-users even though they constitute copyright infringement Filippi and Gracz (2013).

harmonising the rules, to safeguard the investment of database makers and to ensure that the legitimate interests of users of information contained in databases were secured (European Commission, 2005).

Since its adoption by means of the Database directive, the *sui generis* right has received much criticism, including some negative feedback from the CJEU. In Case C-203/02 The British Horseracing Board Ltd and Others v William Hill Organization Ltd[68] the Court explained the essence of the right and limited its scope by holding that "…the investment in creating the materials that made up the contents of a database was to be disregarded and only the investment in collecting them in the database counted." However, as the database right only protects non-substantial parts of a database from copying if such copying conflicts with the normal exploitation of the database or unreasonably prejudices the legitimate interests of the database, the protection may not be as strong as is often assumed (Janssen & Dumortier, 2007, p. 238).

Those who intend to reuse a (complete) database should be aware of potential *sui generis* rights. To avoid an IP law breach, they should abstain from unauthorized reuse. There are two elements of infringement of a database right: first, its extraction and second, its re-utilization. To claim an infringement, it has to affect a substantial part of the database on a one-off basis or repeatedly and systematically of insubstantial parts (Kemp, 2014, p. 16). Those reusers who make use of databases created outside the EU should bear in mind that even if a database right does not exist in the country from which the data originates, there still can be other legal mechanisms in place that restrict free exploitation of databases.

In addition, specific private arrangements can be made between those responsible for the compilation of the database, and the researchers/reusers seeking to utilise the data contained within the database. As RECODE; (RECODE, 2014, p. 15 researchers observed in the interview with a legal expert from the health sector, most often a project officer sets up transfer agreements and decides on which law should be in place. Usually this is the law of the state where the database is held. However, if a database is stored in a cloud, the geographical applicability of laws is more difficult to be established.

3.2.5.1.3 Trade secrets and confidentiality

Contrary to the US federal legal system[69], there is no legislation on the EU level yet that focuses specifically on trade secrets. However, in November 2013, the Commission proposed a draft directive that would align existing laws against the misappropriation of trade secrets across the EU[70]. A new

[68] [2004] ECR II-2905
[69] In the United States, a federal source of law, the Uniform Trade Secrets Act, played a vital role in harmonizing the legal protection of trade secrets across the different U.S. states. In the EU this development has just begun. See for example Czapracka (2007).
[70] http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information#summary-of-the-directive

framework for the protection of trade secretes was confirmed by the Council in 2014 and has been recently moved to the Parliament to continue the regular legislative procedure.

Until the common EU approach is adopted, the area remains regulated by the applicable international and national legal acts. At the international level, trade secrets are addressed by the fundamental agreement in the area of IP law – TRIPS. Article 39 (2) mandates legal protection for breach of confidence or misuse of trade secrets. The protection only applies to information that is secret, that has commercial value because it is secret and that has been subject to reasonable steps to keep it secret (TRIPS, Art. 39). Although all MSs of the EU are also members of the World Trade Organization (WTO) (and therefore TRIPS signatories), their approach to the implementation of Article 39 differs greatly. Some MSs have defined the notion of a trade secret in one of their national legal acts (as part of civil, criminal or commercial law), whereas in other states it has been only introduced as a judicial concept (European Commission, 2013b).

In UK law, the basic IPR's protection of trade secrets is strengthened by the law of confidentiality, a principle of English legal tradition that provides legal recourse against those who disclose secrets to others in confidence and where the recipient subsequently divulges that information to someone else (Drukker Solicitors, 2014). Unlike copyrights and database rights, which are formal remedies protecting information in the way in which it is displayed, the law of confidence can protect the substance of the information itself (Kemp, 2013).

The presence of trade secrets in a dataset has similar consequences for a data reuser as the existence of copyrights or *sui generis* rights. Data reusers should refrain from reusing the data that contains trade secrets and other confidential information unless they have received approval from the beneficiary of the protection or there exists an exception to the general rule.[71]

3.2.5.2 Law of contracts (rights in personam) & licensing

Not harmonized area

In the literature the discussion on licensing is normally linked to the discussion on IP rights. However, by its nature, a licence is no more than a contract, although it has some typical features that cannot be seen in other agreements. For the purposes of this deliverable the discussion on licensing and on contracts are merged and separated from the section on IP law. Our reasoning is justified by the fact that contractual rights are technically entirely disconnected from IPRs and have different legal nature.

[71] RECODE project report describes the example of Tamiflu, a drug manufactured by Roche, a pharma giant. In that case Roche was able to block access to research (trial) data about Tamiflu using trade secret protection.

While IP law imposes obligations on all the parties regardless of their relations *(erga omnes*, *in rem)*[72], contractual duty only applies between the parties of a contract (*intra partes*, *in personam)*[73].

According to Kemp (2014), the key issues in big data contracting include a compromise on the ownership of the rights derived from the data, warranties of compliance with laws and regulations, duration of the relationship and risk allocation. Also critical to address are the issues related to data privacy and security. In addition, most commercial agreements have some form of *limitation of liability* — a provision designed to limit the type and extent of damages to which the contracting parties may be exposed. It is not uncommon to see these provisions disclaim a party's liability for all consequential damages (e.g., lost profits, harm to the reputation of the business) and limit all other liability to some fraction of the fees paid (Kalyvas & Overly, 2014, p. 27).

All these issues have to be determined by data reusers when they enter in a written contract with their data sources. The type of the agreement will, however, depend on the fact how this specific data fits into their business plans. If they want to acquire ownership of the data, they use a data supply contract. If they want to use the data for a defined time period, they can do it under a data licensing agreement. If they want to determine the scope of the right to resell the source's product using the data broker's brand, they enter in a data reseller agreement (2014, p. 16).

In case of data reuse in the public sector, contractual liability is a decisive factor. Public administrative bodies are held responsible if the data is altered, erroneous, not updated, and not suitable to be used for particular purposes or misused, causing damage or loss to other people. As public administrative bodies believe that their data is of insufficient quality to let any third parties use them (unless additional investments are made), the threat of contractual liability often defers reuse of data. The typical response to this concern is either to refuse to make the data available or to impose strict licensing conditions and liability waivers, in order to maintain control over what happens with the data once they are out of their hands (Dulong de Rosnay & Janssen, 2014, p. 6).

Although we have shown that issues related to big data contracting go beyond an agreement on licensing, licensing remains one of the main topics to discuss in relation to big data reuse. A licence is a permission to do something. By licensing, the holder of the rights, regarding, for instance, a symbol, a solution to the technical problem, a work or data, rather than engaging in the manufacture of all the articles (be they bottles, brakes or e-books), authorizes a third party to engage in the production and sale of the same (Ricolfi, n.d.). With regard to data reuse, a licence is nothing more than a contract between a licensor and licensee that defines the scope of activities a licensee may engage in with regard to the licensed database e.g., use the data solely for internal use, distribute limited segments to others,

[72] See p. 32.
[73] *Jura in personam* are rights primarily available against specific persons. Black's Law Dictionary (1910).

combine the database with other data, etc. (Kalyvas & Overly, 2014, p. 91). Terms and conditions or other forms of authorization may be deemed as equivalent of or as containing a licence (LAPSI, n.d.b).

Under the traditional model, the licence granted to the database was generally limited in scope; for example, to a defined set of data or for certain purposes. The licensee generally had a clear understanding of the data being made available to them and what they could do with it. However, it is not as straightforward with respect to big data. The data may consist of data that has been generated by the licensor itself, collected from users and other third parties, licensed from third parties, scraped from the Internet and/or obtained via various social media tools (Kalyvas & Overly, 2014, p. 92). Given the heterogeneity and the complexity of big data, it is of utmost importance that the licence agreement addresses all (or at least the majority of) possibly disputable issues. The most notable ones are explained below.

First, it is necessary to agree on how the IP rights contained in the licensed dataset will be tackled. For example, it is necessary to reflect on the database right and, in the jurisdictions where no database right exists, on applicable copyright protection laws (Kalyvas & Overly, 2014, p. 93). Furthermore, the (future) ownership of data has to be precisely defined in advance. In case of data reuse for the purpose of data mining, for instance, the question will arise regarding whether the licensee or the licensor will claim ownership of derivative works (Kalyvas & Overly, 2014, p. 95).

The licence grant is one of the most significant provisions in the licence agreement (Kalyvas & Overly, 2014, p. 97). First, it determines whether there is a potential for the licensee to benefit from some significant competitive advantage. This will only be the case if the data provided to a licensee will not be provided to his competitors. Second, the provision on the licence grant is important to limit the scope of data reuse and thus prevent unauthorized use of third-party rights such as IP or personal data. However, this will not always be an easy task. When combining sets of disparate data, resulting in the aggregation of data into larger datasets, it becomes increasingly difficult to determine whether the use of such data is within the scope of the applicable original consent the data was collected under and the parties may be subject to potential third-party claims of infringements (Kalyvas & Overly, 2014, p. 99). The parties will also need to agree on the warranties, indemnifications and liability. Big data specifics also influence the nature of these provisions. Given the nature of big data as explained above, its unanticipated uses and the associated risks, most licensors are hesitant or refuse to offer any form of indemnification to a licence (Kalyvas & Overly, 2014, p. 112), although this provision is a standard part of a licence agreement.[74]

Licences proved to be a useful instrument in opening up government data as well as data from private parties. In addition to custom-made licenses, standard licences have gained popularity as a facilitator of data reuse. Well-known examples are Creative Commons (CC), Creative Commons Zero (CC0), open

[74] See for example Pegnberg & Geiddler (2003).

government data, etc.[75] Standard licenses are allegedly more beneficial to users than custom-made alternatives. Apart from the benefits of enhanced organizational efficiency and cost saving, the use of standard licensing terms can lead to greater interoperability of data as well as increased user awareness of the licence terms, thereby enabling better compliance (Korn & Oppenheim, 2011, p. 4).

There have been divergent views regarding the benefits of licensing and reusing open data. In today's Europe, it is a political consensus that open data is advantageous, since by using it, it is easier to break-up, recombine and reuse data in an (ideally) profitable way. Therefore, the economic goal is to give people the right incentives to make their data open and for open data to be easily usable and reusable. In this regard, licensing is important because it reduces uncertainty. Without a licence you do not know where you, stand as a user and whether you are allowed to use the data or to give it to others (Pollock, 2009). On the other hand, RECODE (2014) shares different insights. An analysis of licensing agreements in the geospatial sector has shown that these agreements actually limit further data reuse. In most of the cases the agreements contain a clause that prevents free access to data or its reuse in a purpose non-compatible to the one stated in the agreement. Because of that a number of applications not related to this initial purpose are unable to fully utilise the acquired data.

3.2.5.3 Competition laws

EU COMPETITION LAW	
Primary EU law	Treaty on the Functioning of the EU, Articles 101-106
Secondary EU law	Regulation 1/2003 (modernizes the EU competition law in particular the rules on its enforcement) Regulation 139/2004 (relates to mergers' control)

As a general proposition competition law consists of rules that are intended to protect the process of competition in order to maximise consumer welfare (Whish & Bailey, 2012, p. 1). Competition law is concerned with practices that are harmful to the competitive process, in particular with anti-competitive agreements, abusive behaviour by a monopolist or a dominant firm, mergers and public restrictions of competition (Whish & Bailey, 2012, p. 3). Competition has gained central importance in the EU as one of the most powerful tools the authorities have to restore consumer's welfare (Whish & Bailey, 2012, p. 19). As ex-commissioner Neelie Kroes put it in her speech at the European Consumer and Competition Day in London in 2005, the EU competition law's aim is simple, "To protect competition in the market as a means of enhancing consumer welfare and ensuring an efficient allocation of resources. An effects-

[75] Detailed information about each licence available at: http://creativecommons.org/about/cc0.

based approach, grounded in solid economics, ensures that citizens enjoy the benefits of a competitive, dynamic market economy."

Competition law settles the conditions for a free and unrestricted access to market and this should also be the case on the market of (big, personal) data[76]. With the increasing importance of big data, its legal impacts have been progressively discussed in the academia, in particular in relation to personal data. Yet privacy and data protection are by no means the only aspect of data regulation. National and EU competition authorities have over the last five or so years been showing increasing interest in analyzing through the lens of competition law data in a number of sectors, particularly financial market data (Kemp, 2014).

The regulatory framework in competition law is very complex and a comprehensive overview of all possible implications would require more detailed research of all the aspects of big data activities. We will limit our analysis to the general EU framework and up-to-date research efforts, with the focus on the interplay between competition law and data protection law.

3.2.5.3.1 Big data, competitiveness and privacy

The importance of (personal) data for the competition on the digital market has been assessed very few times. In the Google/DoubleClick case[77] the EC analyzed whether the mere combination of DoubleClick's assets with Google's assets, in particular the databases that both companies have or could develop based on customer online behavior, could allow the merged entity to achieve a position that could not be replicated by its competitors. [78]

The commission also reviewed the case of a merger between TomTom/Tele Atlas.[79] The business goal of that merger was to enable TomTom re-using (integrating) and selling the information acquired from the new business partner Tele Atlas (the merged company).[80] TomTom and TeleAtlas tried to defend the merger with an efficiency claim arguing that that data in the form of feedback from TomTom's large customer base would allow the merged firm to produce better maps faster.

[76] https://research.bournemouth.ac.uk/wp-content/uploads/2014/02/BS-Borghi-Maurizio.pdf
[77] COMP/M.4731, Commission Decision, C(2008) 927.
[78] Another aspect to be considered by the Commission in its investigation was the interaction between competition law and privacy concerns Brockhoff et al. (2008, p. 59). Interestingly, the Commission pointed out that it had referred *exclusively* to the likelihood that the merger would impede effective competition in the common market. However, it noted that its decision was without prejudice to the merged entity's obligations under the Data Protection directive Brockhoff et al. (2008, p. 60).
[79] Case COMP/M.4854, Commission Decision, (C2008) 1859.
[80] TomTom integrates the navigable digital databases it purchases from Tele Atlas into the navigation software the company produces. The integrated product (software and database) is then either included in the personal navigation devices that TomTom itself sells to end-consumers or is sold to other manufacturers of navigation devices for inclusion in their devices.

In cases where data reuse is subject of a competition law discussion it is important that the authorities understand both the competitive benefits and risks of data-driven strategies. Sometimes, a data-driven merger may provide sufficient scale for smaller rivals to effectively compete, however, at other times, data may be used primarily as an entry barrier (Stucke & Grunes, 2015, p. 4). Both, the Google/DoubleClick case and the TomTom/TeleAtlas case were cleared. Still, the fact that a lengthy and costly procedure was initiated confirms the seriousness of the situation and the likelihood of its negative impact to competitiveness in the EU.

In 2014 the European Data Protection Supervisor (EDPS) hosted a workshop to collect best practices and offer more concrete guidance on possible interfaces between data protection law, consumer law and competition law (European Data Protection Supervisor, 2014a). One of the most lucid statements made during the workshop was that a more serious approach to the role of personal information in competition law would encourage the usage of privacy enhancing services (and add to consumer welfare).[81]

The EDPS acknowledged that big data is nowadays considered an important asset in the same way as IP, goodwill or intellectual capital. big data, which is often personal information, plays the role of a currency for purchasing free services. For example, in the case of cross-sided platforms such as Facebook, data is easily (and freely) gained from the consumers on the one side of the market and then sold to advertisers on the other side (Geradin & Kuschewsky, 2013). The EDPS confirmed there was a market for free services where power is accumulated through control of data. If one of the actors on that market acquires a dominant position, this might result in unwilling consequences such as tying, anticompetitive agreements or exploitation of competitors.[82]

3.2.5.3.2 Opposite views

Contrary to the views discussed in the previous section, some authors refuse to see (big) data as an aspect of competition law. For instance, Tucker and Wellford (2014) argue that antitrust law has a

[81] This is how the EDPS explains the reasoning behind his idea of using competition law to strengthen privacy protection, "If companies are compelled to be transparent about the true value of the personal information which they collect, and market analyses take this value into account as part of competition decision, firms might begin to seek competitive advantages by reducing the periods over which information is retained or by providing a 'clear-my-data' button" European Data Protection Supervisor (2014a, p. 33).

[82] EDPS considers the absence of a clear definition of a primary and a secondary data market a fundamental problem European Data Protection Supervisor (2014a, p. 24). Namely, if there is no market definition, regulators cannot determine a dominant position or monopoly. Consequently, they have no basis to claim an abuse of dominance, which is one of the fundamental competition law breaches. Also, they would experience difficulties with proving anticompetitive agreements or mergers. The fact that online firms have been able to convince some (US-based) courts that there is no product market for *free* services like search and convince agencies to treat online media as just another form of traditional media where the consumer can be ignored, paints a false picture Stucke and Grunes (2015).

limited role to play in the era of big data, going so far as to assert that the acquisition and use of big data by online firms is not the type of conduct captured by the antitrust laws.

Craig Richard (2014) discusses the relevance of personal data for antitrust law. He believes that holding large amounts of personal data and having a dominant position on the relevant market does not per se imply harmful effects. The Commission should ensure these concerns are only taken into account when the use of such data has adverse economic consequences – not adverse consequences for data protection alone. Craig admits that a breach of data protection laws could, in some cases, amount to abuse of a dominant position in the same way as the abuse of the patent system led to breach of the competition law. However, this is a very unlikely occurrence, since not many databases can be unique and exclusive, hence leading to a dominant position on the market. [83]

3.2.5.3.3 Reuse in the public sector and competition law related concerns

As mentioned in section 3.2.4.7, the PSI Directive[84] creates a general obligation for MSs to allow reuse for commercial or non-commercial purposes of their freely accessible PSI unless they are protected by an IPR held by a third party. The right of reuse may be granted without any limitations (using the so-called open data licenses[85]) or may be restricted with some conditions.[86] When such conditions are imposed, public sector bodies must comply with competition law and should not discriminate other users (Article 8(2) of the PSI Directive).

The PSI Directive strongly advocates free availability of PSI. In fact, Article 6 of the PSI Directive imposes, as the general rule, marginal cost as the ceiling for fees that may be obtained by the Public Sector Bodies (LAPSI, n.d.a). In this way charges are limited to the extent that enables recovery of specific extra cost but not also cost of collection and production, which is the expense that occurs in the course of their ordinary activities (LAPSI Position paper N.1). In economic terms this means that public bodies cannot make PSI reuse a profitable business.

Lately, incumbent market actors have been submitting claims that this change of policy that was aimed at eliminating competition concerns is in fact illicit, harming their commercial interests. They argue that by opening up their PSI for minimum cost these governments are conducting unfair competition practices or at least inflicting damage (LAPSI, n.d.a).[87] What is more, some public agencies not only sell

[83] Here, he draws a parallel to the Commission's decision in AstraZeneca case (Commission Decision C(2005) 1757 final of 15 June 2005), later tried and upheld by the CJEU.
[84] As amended in 2013.
[85] See the explanation above.
[86] European Union: Step Forward Toward Open Data: The Revision Of The Public Sector Information Directive, Last Updated: November 8 2013, McGuireWoods LLP
[87] Most widely known cases where a private party suits a public body for anti-competitive and unfair exploitation of PSI. See for example Case C-138/11 Compass-databank GmbH v. Republik Österreich [2012] and Koninklijke PostNL B.V. and Cendrid Data consulting B.V. v the State of the Netherlands, Ministry of Infrastructure and Environment [2011].

their raw data but also create their own commercial products by reusing this data. In doing so, they become not only the suppliers of the private sector but also its competitors. Since public agencies often have a monopoly, or at least a dominant market position, distortion of competition becomes a genuine risk. For example, some private actors have invested significant amounts of money to build large road databases. When the government opens a similar database to the general public the users will be more than happy to shift from the commercial provider to a free public service (2014, p. 6). In addition, this behavior seems to obstruct the fundamental idea of PSI re-use laws, which is transferring valuable PSI in reuse to private sector to boost the EU economy. If reuse is already performed by public bodies, the whole concept is corrupted.

Indeed, the underlying idea of competition law is that transactions take place for remuneration reflecting market price. When the public sector bodies give access to PSI without charges both on the wholesale level and in the end-user market the principles of competition law and the PSI Directive may not correspond or even be contrary to each other. However, the question whether this conduct is acceptable does not only depend on the standard competition rules but mainly on the definition of the public task in the national law (LAPSI, n.d.a). The analysis of competition law in case of PSI re-use thus necessarily requires a thorough understanding of public sector and national specifics.

3.2.5.4 Consumer protection

EU CONSUMER PROTECTION LAW	
Primary EU law	CFR 38 TFEU Arts. 12 and 169
Secondary EU law	Directive 93/13/EEC (unfair contract terms) Directive 98/6/EC (price indication) Council Directive 2005/29/EC (unfair commercial practices) Directive 2006/114/EC (misleading advertising) Regulation 2006/2004 (cooperation between authorities) Directive 2011/83/EU (Consumer Rights)

As the EU data protection supervisor made clear in his preliminary opinion (European Data Protection Supervisor, 2014b), consumer protection law plays a visible role in the data-driven economy in particular in ensuring transparency and accuracy of information. The supervisor predicts that the scope for abuse of market dominance and harm to the consumer through refusal of access to personal information and

opaque or misleading privacy policies may justify a new concept of consumer harm for competition enforcement in digital economy.[88]

The UK regulator for markets and competition (CMA) has already embraced this position. In June 2015 it published a comprehensive opinion on commercial use of consumer data listing a number of business practices that are arguably disputable under consumer protection law. For example, according to CMA misrepresenting the privacy, security, or confidentiality of users' information – which could still be deceptive, even if the privacy policy or other small print is factually correct (for example, the consumer is told that data is collected in order to complete a purchase) – violate the provision of fairness set down in the EU and UK national legislation.[89]

Data reusers are bound to comply with data protection law, but in reality they often walk on the edge of law. The fact that behaviour is regulated by both data protection and consumer protection rules means an additional safeguard for data subjects and hopefully more transparency in data reuse.

3.3 Beyond the dichotomy of public and private law

3.3.1 Holistic approach

As shown in section 2, data use, data transfers and data reuse are regulated by different legal acts. We have identified more than ten legal areas that data reusers should be aware of.

However, the legal framework for data reuse can also be structured differently. In the Netherlands, there was a legislative proposal for a 'general data law' presented by the social democrats (PvdA). The idea behind the proposal is the growing inconsistency and uncertainty in the current legal framework for governing data related issues (Schermer, 2015). The general data law would govern how to protect citizens' privacy on the Internet, clearly define the government responsibilities in fighting cybercrime and make clear rules for businesses in the digital economy.

A more coordinated approach towards regulation of data economy has also been proposed in France. In June, the French National Digital Council released the report "Digital Ambition: a French and European

[88] The Commission in 2012 shared a similar opinion:, "In the current economic context a strong consumer policy is a necessity. Empowering Europe's 500 million consumers will be a key contribution to growth in the European economy. The strategy adopted today aims to empower consumers and build their confidence by giving them the tools to participate actively in the market, to make it work for them, to exercise their power of choice and to have their rights properly enforced. We will do so […] by ensuring that consumer interests are more systematically integrated into EU policies of key economic importance for households." (European Commission Press release, A new European Consumer Agenda – Boosting confidence and growth by putting consumers at the heart of the Single Market, IP/12/491. Retrieved from http://europa.eu/rapid/press-release_IP-12-491_en.htm).

[89] Council directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts and its implementation act(s). Contrary to the Consumer rights directive (and implementing acts) the Directive on unfair terms applies whether or not the consumer pays with money – for example if the product is being provided in exchange for personal data (2015, p. 66).

policy for a digital transition", containing 70 proposals for the future of the digital economy in France and Europe. During the same event, Manuel Valls, the French prime minister, announced that its Government would soon introduce a *Digital Bill* aimed at regulating the use of the Internet, as well as stimulating innovation and fostering growth in the digital economy. The bill is expected to also include measures that would give Internet users greater power over their personal data and to allow *portability* of their history and preferences when switching between service providers and social network (Proust, 2015).

A general data law is indeed an attractive idea. However, as Schermer (2015) observes, it is difficult to implement and even harder to enforce.[90] More research is required to find out whether the economy indicates the need for a more unified and simplified legal system and whether this is a sound move.

3.3.2 Sector specific laws

Special regard should be paid to legal barriers that are industry specific. Climate, energy, food, health, transport, security, and social sciences are some of the fields where the processing, analysis and integration of large amounts of data play a growing role, such as the analysis of medical data, the decentralized supply with renewable energies or the optimization of traffic flow in large cities.[91,92] The fact that many of these areas are highly regulated has a big impact on data reuse. For example, the pharmaceutical and the medical devices industry are recognized as two of the most strictly regulated businesses. In addition to general data protection laws, several stringent industry-specific regulations apply to the collection and processing of personal data, in the production process and usage of medical devices and medicines. This is particularly true for clinical trials and investigations, for instance, where substantial amounts of patient data are collected or where customers' data are processed as part of marketing activities (Ursic, 2015). If these restrictions are not clearly imposed and if there are no guidelines, this may create risks and defer the innovation in terms of data reuse. The economic losses may affect all the stakeholders – big international manufacturers as well as citizens seeking better medical devices or medicines.

In later deliverables we will address these industry specific issues in more details, either through an analysis of an industry regulation or through an analysis of regulation pertaining to specific datasets.

[90] "Unfortunately, the proposed solution to this problem, a general data law, is fundamentally flawed. The parliamentarians want to regulate data much the same way as utilities (e.g. electricity, gas, water). What this approach fails to take into account is that legislation for utilities is only concerned with production and transport, not with actual use. But in the context of data precisely this is highly relevant. Another issue is that the data law is to be a Dutch law. This disregards the fact that in today's connected world data flows across borders almost by definition."

[91] The importance of these fields has also been emphasized in the EU's research and innovation framework program Horizon 2020.

[92] http://www.big-data-europe.eu/

3.4 Summary and conclusions

The table below summarizes the findings from the legal analysis. The first column lists different legal areas, the second column describes possible impacts the rules might have to data reuse and its interdependencies, and the third column indicates whether the legislation is typically a barrier to data reuse or its enabler.

Data reuse and EU law		
Public laws		
Data protection law	Every data reuse activity needs a legal basis and every reuser has to adhere to a number of legal principles/conditions e.g. data minimisation, purpose specification, obligations related to data subject rights.	*Most often data protection law acts as a **barrier** (aims at minimizing the amount of data that can be processed, prevents limitless retention, inflicts additional obligations on data reusers etc). However, in specific cases data protection law can also act as an **enabler**, for example through the principle of data quality (an analysis of a more precise dataset would give more reliable results).*
Data Privacy Anti-discrimination Human rights	In certain cases data reuse must be blocked in order to protect fundamental rights such as privacy, non-discrimination etc. (for instance, when profiling is based on discriminatory characteristics).	*Protection of human rights represents an absolute **barrier** to data reuse. When there is a conflict between free data reuse and fundamental rights protection, it is only possible to enhance the latter if data reuse is limited to a certain degree or not applied at all.*
Data localisation	Data localisation provisions restrict data reuse to a limited territory, normally to a territory of a nation state.	*Data localisation is an indirect **barrier** to data reuse. It prevents international data transfers and therefore limits exchange and reuse of data on a global scale.*

Data retention	Data retention provisions encourage data collection and retention by defining a time limit until which the data has to be stored and must not be deleted.	*By ensuring the data is available for additional analyses and examination in the future, data retention law acts as an **enabler**.*
Cybersecurity	Imposes additional network security obligations.	*Cybersecurity can be both, a barrier or an enabler to data reuse. If the requirements are too burdensome, then they deter processors from keeping the data and reusing it (**barrier**). However, by stimulating security those requirements can create a more trusted and secure environment that actually encourages data reuse (**enabler**).*
National security	Prohibits data reuse if data contains sensitive national security related information.	*In those specific cases national security means a **barrier** to data reuse.*
Private law		
IPRs law	Protects the data that contains IPRs and restricts data reuse.	*IPRs are **barriers** to data reuse when they limit data reusers to fully exploit datasets. They are **enablers** when they guarantee better legal protection and thus encourage authors to share their (IP protected) data.*
Competition law	Prohibits data reuse if the company that processes the data has a dominant position on the market. Encourages data reuse on the market as a whole, allowing also non-dominant firms to benefit from big data.	*It is a **barrier** when it limits a data owner that has a dominant position in the market. It is an **enabler** when it encourages fair competition.*

Contracts & licencing	Regulate contractual agreements between reusers and suppliers.	*Depending on actual contract terms, the contractual provisions can act either as a **barrier** or as an **enabler**.*
Consumer protection	Regulates data reuse to benefit consumers.	*When data reusers reuse data to customize their marketing strategies and reach out to more clients, consumer protection law blocks those business conducts that are based on unfair practices (e.g. broadly defined purposes of data processing to enable easy contact with the costumers). On the other hand, data reuse can also be advantageous to consumers, therefore acting as an enabler, when it (justly) helps achieve more precise and targeted advertising.*

The table above makes it evident that many types of legislation are barriers rather than enablers for data reuse. In the EuDEco project/next deliverables, we will investigate how more data reuse may take place within the existing legal frameworks and how the legal framework may/should be modified in order to foster the exchange of data assets. Also, we will elaborate on what is the role of legal certainty in that process and how can the law help set up pilots of a self-sustaining data market.

4 Framework conditions: A socio-economic perspective

4.1 Introduction

From a socio-economic perspective, with respect to framework conditions, EuDEco focuses on the principles behind social and economic realities in the data economy. Accordingly, particular attention is put on the behaviour of agents and the market mechanism.

This means that with respect to the social dimension, principles of user behaviour as well as characteristics of interaction and collaboration will be investigated. Interaction and collaboration is of interest for EuDEco on the level of individuals as well as on the level of organisations. The emergence of value networks in the data economy, where organisations and individuals interact and collaborate with each other to benefit the entire group, receive particular attention.

With respect to the economic dimension, EuDEco focuses on principles of business models, deployment models, data pricing (e.g., free vs. paid data access) and data marketplaces. With respect to data marketplaces, for instance, data-related criteria such as accuracy, comprehensiveness, currency/timeliness or granularity are of particular relevance.

4.2 Societal perspective

What are the societal framework conditions that support and/or hamper large-scale data reuse in Europe? Society and the digital world are intertwined and influence each other, and it is also the case for so-called *big data*. Societal stakeholders encompass businesses, government, local authorities, researchers, civil society and citizens. While nearly all of them are data producers of some kind, only some of them are also data owners and data analysers. One objective of EuDEco is to identify and responsibly overcome societal and cultural barriers to a big data economy in Europe.

The first part of this section (section 4.2.1) explains why big data is a socio-technical phenomenon. The second part (section 4.2.2) highlights the level of awareness and knowledge among European stakeholders about the big data situation. The aim is to understand whether the expectations and fears of socio-political actors may have an enabling or limiting impact on the development of the data economy. Afterwards, the specific case of personal information voluntary disclosed by the individuals themselves is analysed through the prism of psychology (section 4.2.3). Conversely, other individuals are implementing strategies to prevent the analysis of their online digital traces (section 4.2.4). The last parts analyses how social and cultural barriers in Europe can be a hindrance to the model deployment in Europe, either by limiting data sharing (section 4.2.5) or by making data interpretation more complex (section 4.2.6).

4.2.1 Big data is a socio-technical phenomenon

Like other technological innovations, big data is produced by the society and, in return, may impact it. Big data is thus called a socio-technical phenomenon (boyd & Crawford, 2012), which is likely to impact the way people communicate, live, do business and science; and is fed by all these activities.

Big data and its analysis are cross-sectorial. Massive amounts of data are generated every day, for example in the financial sector, subsequent to bank transfer, stock exchange, purchases with credit cards, etc. Big data analysis sometimes becomes a paid service – called Big Data as a service (BDaaS) –, which enables, among others, companies to fine-tune business strategy and to target advertising. Data can also be used to improve science, public services, and security (e.g., for charting epidemics, improving traffic flow, forecasting weather, etc.). Precise capabilities of data analysis are yet uncertain.

Increasing numbers of connected actions can be tracked, measured and linked to real people, machines, dates and places. Even the vocabulary is changing quickly. The Internet of Things is becoming the Internet of Everything (IoE); a term recently coined by Cisco (Bradley, Reberger, Dixit, & Gupta, 2013).

The IoE is meant to be broader than the Internet of Things (i.e. data created by connected objects) as it also includes data on processes and people.

It is therefore essential to acknowledge that the data ecosystem is wholly embedded in societal framework conditions, which impact and are impacted by the rise of the big data phenomenon.

4.2.2 Perceptions of big data issues

This sub-section presents an analysis of the big data perception by public and private stakeholders and its potential impact on the data economy model. The EuDEco project has to take into account big data perceptions in Europe in order to understand whether it may have a facilitating or constraining impact on data reuse and on the data economy.

4.2.2.1 What is the level of awareness on big data issues?

Even though many citizens are not familiar with the term *big data*, introduced in the late 1990s, and what it means – why and how big amounts of data are collected, centralized, stored, analysed and reused – a general awareness on related challenges and issues can be observed. The press can serve as an indicator of the European population's interest for big data in general, and in for the data economy in particular. A wave of articles covering this topic is regularly published both in mainstream media and in specialized journals (European Physical Journal, L'Usine Nouvelle, etc.). Debates have been fuelled by notorious cases related to negative uses of big data, such as espionage scandals (global surveillance disclosures, since 2012) or to positive uses, like data used for tracking epidemics (attempt to monitor the expansion of Ebola and flu).

In addition, several peer-reviewed journals addressing Big Data & Society topics have been recently created, e.g., Big Data & Society (BD&S), a SAGE open access peer-reviewed scholarly journal, created in April 2014; and the Journal of Big Data, a Springer open journal, created in June 2014.

Along with the expanding number of publications, new fields of study related to big data are emerging. One of them is Critical Data Studies (CDS), "a growing field of research that focuses on the unique theoretical, ethical, and epistemological challenges posed by Big Data" [93]. The CDS claims that big data should be analysed as a phenomenon that is not ideologically neutral. Other authors (e.g., Kitchin, 2014) consider that existing fields of study, in all fields of science, are subject to a paradigm shift. Kitchin argues that the recent availability of big data, combined with data analytics, is creating a new form of empirical analysis and data-driven science. This may change the way researchers make sense of culture, history, economy and society.

In conclusion, a wide range of stakeholders, including the general public, companies and researchers show interest in (big) data reuse and analysis. One explanation is that, because the analysis of big data is

[93] Special Theme on Critical Data Studies, Big Data & Society Journal, 22 June 2015

tending towards omniscience ability, it raises questions interesting the whole society. In this context, *de facto*, most societal stakeholders are – and feel – concerned with major questions arising from the big data boom.

4.2.2.2 Facing the big data dilemma

Analysing citizens' perception of the data economy and data reuse is of particular interest because it highlights possible barriers to the implementation of a broadly accepted data economy model. Societal actors are facing a dilemma between (1) the opportunity to benefit from the use of innovative devices and services, based on data collection and analysis, and (2) possible consequences, particularly in terms of preserving one's privacy and business confidentiality.

An in-depth analysis of literature on society and data issues reveals that a significant part of the articles is focused on the ambivalence between pros and cons about big data (with variations such as attraction and fears, benefits and risks, utopian and dystopian futures, etc.). Researchers highlight that people are facing a dilemma between on the one hand attraction for incentive new tech and services and, on the other, privacy fears (Bollier, 2010; Kitchin, 2013; Kshetri, 2014).

A general trend is a growing success, among all societal actors, of technologies that generate big data, sometimes without the user's knowledge. These technologies provide benefits of different kinds: economic advantage (possibilities to communicate, to do business and to use various tools), cultural advantage (opportunities to learn, to get information and to express oneself on the Internet) and entertainment (games, and, more broadly, the growing *gamification* of online content).

One example of disruptive data measurement is telemetry, which is used in the video game industry to capture information about online game activities. Actions undertaken by actors while navigating a game are recorded and analysed. According to some experts (Franks, 2012), this new technology "takes a video game producer well beyond simply knowing how many customers bought a game and perhaps how many hours the games are played. Telemetry data makes it possible for game producers to know intimate details about how customers actually play and interact with the games they've created. The amount of data captured can be huge, and the video game industry is just starting to analyse this data in earnest".

Overall, peoples' taste for new devices and applications – generating big data – is becoming a cultural phenomenon, reinforced by marketing strategies. Volume and diversity of collected data are thus increasing as a consequence of both technological improvements and society's appeal for innovative tech and tools/services/games they provide.

Big data issues entail big challenges and uncertainty, resulting in ideological bias such as the construction of utopian or dystopian forecasts on the use of big data. These passionate debates are sometimes based on literary and philosophical works that have made up the fabric of Western

collective psyche. Data centralization and surveillance are longstanding issues, which can bring either progress or danger in the future. In literature, the famous *Big Brother* (Orwell, 1949) has highlighted a pessimistic view of this issue half a century ago: the figure of a totalitarian state wherein every citizen is under constant surveillance by the authorities. In architecture, the digital world is sometimes compared to the Panopticon, a fictional building invented by Jeremy Bentham in the late 18th Century. The objective of the Panopticon structure is to allow a watchman in a central tower to observe all the prisoners locked in individual cells around the tower. The prisoners can't know when they are observed. This situation creates a sense of invisible omniscience and impacts the way inmates act, as they have to constantly control their own behaviour.

These popular references, about a stakeholder (e.g., a state) having uncontrolled access to people's information, have contributed to shape a common culture in Europe about potential negative impacts of a big data economy. This general background is called a "mythology that provokes extensive utopian and dystopian rhetoric" about the rise of big data (boyd & Crawford, 2012). Conversely, utopian discourses on big data benefits can nowadays be found on websites of business companies, such as (among others) the Utopia Global Inc.[94], a company providing big-data-based services and which motto is *Perfect Data, Perfectly Possible*.

The EuDEco model will obviously go beyond those radical, simplistic visions, while taking into account the hopes and fears they are symptoms of.

4.2.3 Psychology of mass self-communication

The origin of data used for economic purposes can be analyzed in order to understand the reasons behind their disclosure. Psychology can provide insights into the causes leading to the specific case of personal information voluntary disclosed by the individuals themselves. Indeed, people voluntarily stream on the Web various amount of personal information due to an impression of freedom, and sometimes anonymity. The will to keep the Internet out of governments and companies' control has been expressed, e.g. in the Declaration of the Independence of the Cyberspace in 1996 (Barlow, 1996). Self-production of massive private data can also be generated following the will to humanize communication on Internet, for example by disclosing information on personal background and by posting personal pictures (Price, Adam, & Nuseibeh, 2005). Putting pictures and information of oneself in online networks – sometimes thought to be private – can create a sense of closeness with others, even at risk of losing individuals' privacy. This trend is reinforced by narcissistic tendencies (Ryan & Xenos, 2011). This culture of promoting apparent freedom and self-transparency is particularly manifest in North America. Among the negative consequences, there is a risk of unclear distinction between public sphere, work sphere and private sphere.

[94] http://www.utopiainc.com

Massive online data production by individuals is raising an ethical question as the disruptive big data phenomenon "has surpassed the capacity of the average consumer to understand his or her actions and their knock-on effects" (Zwitter, 2014). The author of the article, Andrej Zwitter, argues that, in the big data age, individuals don't have the necessary knowledge to estimate some of the consequences of the digital traces they have left behind. Ethical responsibilities should then be reconsidered.

4.2.4 Strategies for limiting big data analysis

The preservation of people's privacy and business' confidentiality is based on a set of legal, organizational, technical and behavioural means. These means are used to protect people and organizations (businesses, institutions, etc.) from unintended and harmful data reuse. Legal aspects include the necessity of constantly adapting European and national laws to technical innovations (see section 3). Organizational means are specific internal rules and procedures aiming to protect data of the organizations (businesses, institutions, etc.) from unauthorized disclosure and data modification. Technical means are digital security techniques (like cryptography) developed to ensure data protection (see section 5). Finally, behavioural means are best practices and strategies for limiting the dissemination of personal/confidential data, or for making their interpretation more complex, thus reducing surveillance risks. However, it is important to differentiate *deep data* (about a few) and *surface data* (about the many), the last one presenting lower risks of breaching privacy/confidentiality (Manovich, 2011).

One example of strategy for making data interpretation more complex on the Internet is the Data Obfuscation (DO). The obfuscation is a strategy for protecting privacy, preserving confidentiality and managing online reputation, by flooding large amounts of highly visible information that will hide the undesirable one (Parameswaran, 2006). The term *obfuscation* came from the Latin *obfuscatio,* which means *to darken*. This strategy finds its roots in a context in which online information can be inaccurate, embarrassing and even defamatory, and rectifying or deleting this undesirable information is often difficult. A Firefox browser extension, TrackMeNot, is based on obfuscation operations: it creates numerous random queries in Google search engine in order to overwhelm real queries (and thus real area of interest of the Internet user) under false queries. Other strategies exist in attempt to cover the trail of digital breadcrumbs left by people' daily activities. However, these circumvention strategies remain marginally used by the general public and the implications of big data analysis are still unclear.

A right to be forgotten on the Internet is becoming a reality in the European data protection law. The right to be forgotten would be a solution to "the titanic clash we're about to witness between privacy and free speech on the Internet" (Rosen, 2012).

4.2.5 Creating a data culture in European companies, public institutions and society

Creating a data culture in companies and public institutions – that is so say promoting data reuse, or even more, open data – may be a key criterion for enabling an extensive development of a data

economy. This issue was debated for example during the conference Building the Big Data Economy: How Data Innovation Is Driving Performance, hold in June 2015, Washington D.C., where panellists discussed best practices for building positive cultural change in data reuse and ultimately for driving business benefits (Clark, 2015).

Who is currently benefiting from the big data economy? A survey conducted in France by Ernst & Young in November 2014 revealed that despite benefiting from a largely positive perception in French companies, only a few of them have developed big data activities (EY, 2014). The world big data activities are currently highly concentrated, especially in the big four (American) giants of the Internet: Google, Amazon, Facebook and Apple, also called the GAFA (Rampini, 2015). In 2013, the revenue of the GAFA was $316 billion, which was similar to the GDP of Denmark ($330 billion) (Fabernovel, 2014). This raises the question of how to ensure that the emerging data economy will benefit to European companies and, more broadly, to all of society.

The promotion of a data culture in European companies, public institutions and society, is a key enabler for the development of a self-sustaining data economy. Emma Muckersie, a British consultant in data analytics highlighted the lack of data culture in Europe, "Unlike our friends in America, who have embraced big data with open arms, the British public remain rather sceptical. We are much more reluctant to allow companies access to our data, worrying about how it will be used in the future. So whilst we may have the technical know-how to analyse big data sets, we are still being held back by the fact that the public do not want to share their data" (Muckersie, 2013). One part of the solution would be to restore public faith in data sharing by focusing on the cases in which big data has been used for the public good (to aid medical research, to improve customer service). Another part of the solution would be to educate all organisations about best practices for collecting, sharing and analysing data and by making clear the way personal data will be used.

4.2.6 Barriers of data incompleteness and skills gap

Another barrier to the data economy and data reuse is the difficulty to analyse data that is inherently diverse and non-exhaustive. Those difficulties are increased by the diversity of cultures, languages and norms in Europe – for example, there are 24 official and working languages in the EU, and many more spoken throughout Europe. It is also important to note that in many cases and for many purposes data may be non-existent or non-exhaustive. For example, around 20% of EU individuals have never used the Internet and 50% have never used social media (Seybert & Reinecke, 2013).

Finally, extracting value and relevant information from large datasets requires interpretation capacities. In 2012, a study predicted that in the UK alone, the number of specialist big data staff working in larger firms would increase by more than 240% over the next five years (e-skills UK, 2013). The need of new skills for analysing big data was also outlined by Hal Varian, Chief Economist at Google, in an interview in 2009, who predicted, in a humorous way, "I keep saying that the sexy job in the next ten years will be

statisticians, and I'm not kidding". Because having massive amounts of data doesn't necessary create meaning, the data economy would require increased interpretation capacities.

4.3 Economic perspective

What are the economic framework conditions that support and/or hamper large-scale data reuse in Europe? What kind of economic factor is data? What makes an economy a data economy?

Data volumes are exploding and a slowdown is out of sight. Companies, government bodies, academic institutions and citizens have access to more data today than anyone would have imagined one or two decades ago. Traditional data sources such as company databases and applications are now complemented by non-traditional sources such as social media or sensors embedded in physical world devices including mobile devices, smart meters, cars and industrial machines. Simultaneously, an entirely new market of big data technologies and services has emerged over half a decade to help organizations capture and extract value from all the data. The revenue from big data technologies and services, however, is small compared to the value that is expected to result in sectors such as trade, manufacturing, finance and insurance, public administration, and health and social care that now have the tools at their disposal to make innovative use of data to drive high-value business and societal outcomes. Data and technology are simply enablers. It is how society puts data and technology to work that will drive value creation and define the data economy.

The data economy is the currently emerging economy in which participants succeed or fail based on their ability to leverage data and analytics to improve operational efficiencies, to make better tactical and strategic decisions, and to create innovative products and services that meet and exceed customer expectations. Additionally, the data economy is also a connected economy in which partners and competitors alike share data and integrate business processes where the expected benefits for the participants and their customers outweigh the risks of such collaboration. The data economy has already begun to take shape. Companies from the Web sector such as Google, Yahoo and Facebook were among the first to understand the potential benefits of harnessing big data technology and services but business leaders across sectors are finally buying into the transformative power of the connected and intelligent enterprise.

However, to realise the full potential inherent in the data economy, independent use cases and applications of big data must be part of a larger whole. Some expect that entire sectors will operate and markets function all through the intelligent use and sharing of data and that eventually enterprises in various sectors will orchestrate multiple applications to work intelligently together with the goal of optimizing entire operational environments. Without doubt, the data economy is not developing uniformly across the developed world. The development opportunities as well as the pace of development of the data economy are determined by the framework conditions given in a country or a group of countries characterised by harmonised framework conditions such as the EU. From an economic perspective, aspects such as the macroeconomic environment, investments in research and

development, the availability of capital and skilled labour as well as the size and efficiency of the data market are considered framework conditions.

4.3.1 Financing economic activity – the capital market

4.3.1.1 Characteristics and key instruments

Capital markets channel capital between suppliers of capital such as individual investors, institutional investors (e.g., pension funds, mutual funds), government bodies, companies and financial institutions, and users of capital. They are vital to the functioning of an economy, since capital is a critical component for generating economic output. Capital markets differ from money markets in that the former ones are accessed to raise capital for long-term purposes, such as for a merger or acquisition, to expand a line of business or enter into a new business, or for other capital projects. Suppliers of capital generally want the maximum possible return at the lowest possible risk, while users of capital want to raise capital at the lowest possible cost.

The options to raise money on the capital market vary considerably depending on whether an established company wants to expand its business or a startup wants to start a new business. The focus of this section will be on startups. Established companies typically use capital market instruments such as stocks and bonds, or business loans. Regular bank lending as well as some of the instruments relevant for startups describe below are, strictly speaking, not considered as a capital market transactions[95].

Small business loans are usually provided by banks. Lending money from banks allows doing business without giving up control but interest has to be paid based on the banks' banking conditions and risk assessment. Under certain circumstances, banks will ask for collaterals or a detailed business plan including a market study. Data is not very suitable for use as collateral and creating a convincing business plan is difficult, particularly, as the value of data – and in consequence also the return on investment – is very difficult to estimate. Moreover, the principles of the data economy are not yet fully understood and information on competitors and prospects is difficult to get and rather uncertain.

Venture capitalist investors look specifically for start-ups to fund. This option has a lot of money available to offer to startups and plenty of resources to actually help startups succeed. However, there are a few major downsides to this option. Investors typically look for opportunities that are relatively stable, meaning the startup has a strong team of people and already acquired a considerable amount of funding. Moreover, the mentorship which is typical for venture capital requires giving up some control.

Crowdfunding is a newer form of funding startups. Platform such as Kickstarter[96] and Indiegogo[97] are examples of websites where crowdfunding campaigns can be created. Anyone can contribute money

[95] http://www.inc.com/adam-heitzman/5-best-ways-for-funding-a-startup.html
[96] https://www.kickstarter.com/
[97] https://www.indiegogo.com/

toward helping a startup that they believe in. Those giving money will make online pledges with the promise of pre-buying the product, earning some type of reward or giving a donation.

Angel investors work similarly to venture capitalists except they are much smaller in terms of operation, sometimes only one person. Such investors often want a large portion of a startup. This is a very popular options because it allows keeping control over the company, earn mentorship when it's needed, and make money as the company continues to grow. A few popular places to meet angel investors include Small Business Development Centers and platforms such as Gust[98].

An alternative to taking advantage of the capital market or using other funding instruments is *bootstrapping*. Bootstrapping means building a business while seeking to avoid relying on outside investors (Winborg & Landström, 2001).

4.3.1.2 Differences between Europe and the US

According to Collin (2014) venture capital funding in Europe and in the US are two very different things. First, there are more investors with more money in the US than in Europe. Investors in the US are chasing founders as much as founders are chasing them. Figure 2 shows that the funding gap is increasing. In 2014, European companies have been receiving more venture capital than ever before. European startups raised about $7.6 billion last year, a 41 percent leap over 2013. But that was only about one-fifth the amount raised by US technology companies, which secured a combined $37.9 billion in 2014, up more than 30 percent from the previous year. Second, scepticism seems to be more widespread in Europe than in the US. European investors typically request a three-year business plan as a prerequisite to any form of conversation. It seems that the path to profitability has to be already proven. According to Collin (2014), in the US it felt like taking a risk together. Third, investors in the US make decisions faster. The average time to raise funds in Europe is three to six months. In the US, raising funds is possible in a third of the time.

[98] https://gust.com/

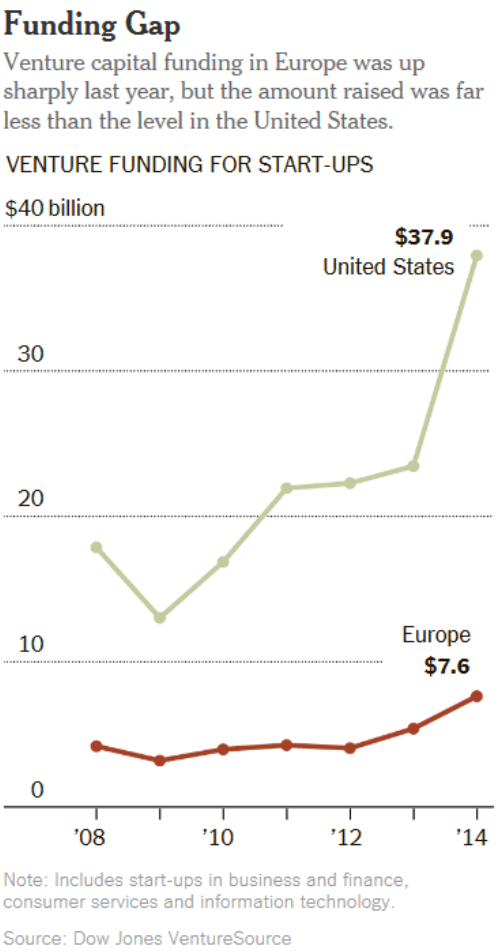

Figure 2: Funding gap between Europe and the US[99]

Scott (2015) highlights that there is a better awareness by US startups about the importance of global expansion. US entrepreneurs once saw international expansion as secondary to domestic growth. But the rise of copycat companies around the world has spurred the Americans to spend internationally before local rivals can become entrenched.

4.3.1.3 SAP Startup Focus – a concrete case

SAP launched Startup Focus[100] as a global program to help promising startups in the big data space develop new applications on SAP HANA and accelerate market traction. SAP HANA[101] is an in-memory,

[99] http://www.nytimes.com/2015/02/14/technology/for-tech-start-ups-in-europe-an-oceanic-divide-in-funding.html?_r=0
[100] http://www.sap.com/corporate-en/about/our-company/innovation/start-up-focus-program.html
[101] http://discover.sap.com/hana/index.html?url_id=banner-us-homepage-row1-full-rs-discovery-july15r2

column-oriented, relational database management system (RDMS) developed and marketed by SAP. SAP Startup Focus provides startups with

- access to the HANA technology platform
- access to technical HANA experts and marketing & sales enablement support
- access to its global customer base to sell market-ready solutions
- the opportunity to pitch to the venture community via the HANA Real Time Fund and SAP Ventures
- contacts with entrepreneurs, partners, investors and thought leaders in the big data space

SAP Startup Focus does not require any equity sharing from participating startups.

4.3.2 Institutional landscape

This section focuses on the attitude towards big data and data reuse of institutions relevant in the context of the data economy. Both public and private institutions receive attention. The attitudes of the institutions as well as the measures they implement have a lasting effect on the data economy as they affect the incentives of and constraints on actors in the data economy (Acemoglu, Johnson, & Robinson, 2005). The institutions determine the future of the data economy in Europe and beyond to a considerable extent.

Big data and data reuse are relatively new topics in the public debate. Only recently, institutions have started making their points of view explicit through the publication of strategies or policies. There is wide consensus that the opportunities that come with the data economy are accompanied by challenges that need to be addressed. Attitudes and measures of institutions have widely been consistent over time and it seems that societal concerns are generally taken into account.

So far, the relevant institutions seem to play their part in the provision of a stable and predictable environment for the data economy. Decision making processes seem to be transparent and the trend to making government data publicly available, in addition, helps to ensure transparency. Institutions participating in the open data movement may even become role models for other actors in the data economy.

While the attitude of public institutions is mostly relevant for the development of the data economy because they define the framework conditions, the attitude of private institutions is relevant because their approach towards corporate social responsibility defines how big data applications look in practice. Acting in a fair and transparent way as well as implementing concepts such as privacy by default and by design will help to raise acceptance of big data and data reuse and benefit from the opportunities it provides both economically and socially.

4.3.2.1 EU level institutions

The EU recognizes big data as a driver of the future European economy and urges that its potential be exploited (NTT DATA, 2015). With respect to the European data economy, a particularly relevant institution is certainly the EC. Efforts of the EC related to the data economy are, by and large, part of the Digital Agenda for Europe (European Commission, 2010). The Digital Agenda is one of the seven pillars of the Europe 2020 Strategy which sets objectives for the growth of the EU by 2020. The main objective of the Digital Agenda is to develop a digital single market. In order to achieve the digital single market the EC is taking several actions. Some of the actions are highly relevant in the context of big data and data reuse.

The EC selected making big data work as one of the key goals to be pursued within the scope of the Digital Agenda[102]. Within the scope of its efforts, the EC:

- develops a big data strategy,
- funds research and innovation activities in the fields of big data and open data (see section 4.3.3),
- develops an open data policy, and
- facilitates use and reuse of publications and data resulting from research (see section 4.3.3).

In terms of big data strategy, the EC drafted the Communication on the data-driven economy (European Commission, 2014c) as a response to the European Council's conclusions of October 2013 (European Council, 2013), which focused on the digital economy, innovation and services as drivers for growth and jobs and called for EU action to provide the right framework conditions for a single market for big data and cloud computing. This Communication describes the features of the data-driven economy of the future and sets out operational conclusions to support and accelerate the transition towards it. It also sets out current and future activities in the field of cloud computing. According to the Communication, a thriving data-driven economy has the following characteristics (European Commission, 2014c, pp. 5–6):

- Availability of good quality, reliable and interoperable datasets and enabling infrastructure
- Improved framework conditions that facilitate value generation from datasets
- A range of application areas where improved big data handling can make a difference

In terms of open data policy, the EC's work focuses on generating value through reuse of PSI. The EC took legislative and non-legislative measures and encourages the creation of open data portals. The EC drafted a Communication on open data presenting measures that it considers suitable to overcome existing barriers and fragmentation across the EU (European Commission, 2011a). The Directive on the reuse of PSI (Directive 2003/98/EC) entered into force in 2003. It was revised by Directive 2013/37/EU which entered into force in mid-2013. It focuses on the economic aspects of re-use of information rather

[102] https://ec.europa.eu/digital-agenda/en/big-data

than on the access of citizens to information. In July 2014, the EC published guidelines to help the MSs transpose the revised rules and to indicate best practice in several fields important for the reuse of PSI (European Commission, 2014b). In additions to legislative measures, the EC set up the PSI Group[103], funds the ePSIplatform[104] and the LAPSI network[105], commissions studies on PSI-related issues, develops an open data portal and contributes to the G8 process on opening up government information. To facilitate access to and resuse of PSI, portal websites on public data have been developed. The EC developed an open data portal site for any type of information held by the EC and other EU institutions and bodies. Moreover, the EC develops a pan-European digital service infrastructure for open data and works with the MSs on data formats and interoperability between open data portal sites. Shadbolt (2010) outlined how an EU portal aggregating national efforts could look like.

Another institution, the Big Data Value Public-Private Partnership (PPP)[106], aims at strengthening the data value chain, in order to allow Europe to play a relevant role in big data in the global market. In order to cooperate in data-related research and innovation, enhance community building around data and to set the grounds for a thriving data-driven economy in Europe, the EC as well as European industry and researchers have teamed up in a PPP. The partnership between industry and researchers forms the Big Data Value Association (BDVA)[107]. A contractual arrangement on the PPP has been signed in October 2014. In order to reach its objectives, the PPP will make use of two major instruments:

- Large lighthouse projects in candidate areas like manufacturing, personalised medicine and energy;
- i-Spaces, environments where infrastructure is made available to bring technology providers and end-users together to identify services, skills, business models and ecosystems in which novel technology and applications can be built.

The objectives of the BDVA are to boost European big data research and innovation as well as to foster a positive perception of big data. It aims at:

- strengthening competitiveness and ensuring industrial leadership of providers and end users of big data technology-based systems and services;
- promoting the widest and best uptake of big data technologies and services for professional and private use;
- establishing the excellence of the science base of creation of value from big data.

[103] https://ec.europa.eu/digital-agenda/en/news/public-sector-information-group-main-page
[104] http://www.epsiplatform.eu/
[105] http://www.lapsi-project.eu/
[106] http://ec.europa.eu/digital-agenda/en/big-data-value-public-private-partnership
[107] http://www.bdva.eu/

Another one is DIGITALEUROPE[108], which represents the digital technology industry in Europe. Its members include some of the world's largest IT, telecommunications and consumer electronics companies as well as European national associations. The institution published a position paper describing measures for an action plan making Europe fit for the data economy (DIGITALEUROPE, 2014). DIGITALEUROPE stresses that while certain risks associated with big data need to be addressed, the focus of an action plan would have to lie on fostering opportunities related to the data economy. The paper acknowledges the efforts the EC has made previously. DIGITALEUROPE considers the following measures particularly relevant to enable the digital economy to thrive:

- Adopt a harmonised framework for data protection
- Protect big data applications from cyber-attacks
- Ensure interoperability through global standards
- Clarify copyright rules to facilitate text and data mining
- Boost the deployment of open data
- Create trust in cross-border data flows
- Fill the data skills gap
- Support investments in infrastructure and research

As DIGITALEUROPE is representing the industry, it does not come as a surprise that it emphasises that a modern framework for data protection should not only create trust but also enable innovations, that regulatory efforts related to cyber security should focus on critical infrastructures or that standards should be voluntary and market driven.

4.3.2.2 National institutions in selected EU MSs and the US

On the national level, the most relevant institutions in the context of the data economy are the national governments.

The UK government in association with the EU considers big data a vital driver of the UK economy and listed it under one of the elements of the *Eight Great Technologies* for the country to remain globally competitive[109]. The UK operates an open data portal[110]. A UK private institution worth mentioning is the Open Data Institute which aims at spreading and developing the principles of open data technology[111].

The French government is also attempting to promote data research and innovation. In April 2013, the government launched the Innovation 2030 Worldwide Challenge under the Innovation framework to fund innovation projects. This contest has seven goals with big data as the seventh goal[112]. France

[108] http://www.digitaleurope.org/
[109] http://www.stfc.ac.uk/research/technologies/the-eight-great-technologies/
[110] http://data.gov.uk/
[111] http://opendatainstitute.org/
[112] http://www.entreprises.gouv.fr/innovation-2030/seven-strategic-goals?language=en-gb

operates an open data portal[113]. With Opendata France[114], there is also an institution which support the regional governments in their approach to make PSI available.

Only recently, a public debate on the technologies used for national security purposes and their meaning for privacy stated[115]. Earlier, the opportunities of big data and data reuse seemed to be in the focus of attention almost exclusively. In mid-2004, a first report was published taking not only big data opportunities but also the preservation of values into account (Executive Office of the President, 2014). The report acknowledges that big data can alter the balance of power between government and citizen, reveal intimate personal details and lead to discriminatory outcomes. The report states that it is possible to ensure innovation while protecting values through law, policy and practices encouraged in the public and private sector. The US operates an open data portal[116].

4.3.3 From research and development to business sophistication

To thrive and to be competitive, a data economy needs an environment that is conducive to innovation activity. Ideally it is supported by public as well as by private institutions. It is essential that investments in research and development are considered worthwhile. This requires not only high-quality scientific research organisations as well as a good cooperation between industry and research but also a suitable legal framework (related aspects of private law such as IPR are discussed in section 3.2.5).

The EC as well as national and regional governments invest significant amounts of money into big data research and development. Competence centres are being installed and projects are funded.

Germany, for instance, puts efforts into a promising project called Smart Data Innovation Lab[117].

The BDVA is a key actor with respect to research and development in Europe.

The market dominance and market power of some additional actors reduces the possibilities to enter the market. This may reduce the innovation potential. Therefore the level of competition needs to increase in some areas by supporting startups. One example of a market equilibrium seems to be the adverting data market. Many actors are acting at low costs. Supply and demand determinates the price which is due to the open market and the large number of data sources.

With the ODINE (Open Data Incubator Europe) project, there is finally an incubator for startups in Europe. The focus of ODINE, however, is on open data companies being founded.

[113] https://www.data.gouv.fr/en/
[114] http://www.opendatafrance.net/
[115] https://www.whitehouse.gov/issues/technology/big-data-review
[116] http://www.data.gov/
[117] http://www.sdil.de/en/

4.3.4 Labour market, education and training

The need for skilled labour is one the main issues in the data economy (OECD, 2006). Availability of appropriately qualified staff requires high-quality education and training. STEM skills (science, technology, engineering and mathematics) are not only relevant for the data economy (Capgemini Consulting, 2013) but for the economy as a whole as studies in Canada (Council of Canadian Academies, 2015), Australia (The Australian Industry Group, 2015), the UK (Bosworth, Lyonette, Wilson, Bayliss, & Fathers, 2013) and the US (Atkinson & Mayo, 2010) show.

As the Wall Street Journal mentions, the inability to find and keep workers with even moderate math and statistics skills is already placing limits on business profitability (Aeppel, 2014). In 2014, 500.000 big data jobs were already existing in the US. It was estimated though that there was a need for many more data scientists (Mandel, 2014). The situation in Europe does not differ considerably from the one in the US. There may even be a competition for talents between the US, Europe and other prospering regions of the world.

It is considered important to differentiate between data scientists with deep expertise in statistics and machine learning and elements of data science that should be provided to students of all levels of education from high schools to universities (Giannotti, Nanni, Rauber, & Thanos, 2014):

- Data scientists: A significant constraint on taking advantage of big data is the shortage of talent, particularly of people with expertise in statistics and machine learning, managers and analysts who know how to use insights from big data effectively, and technicians who develop, implement and maintain the hardware and software tools needed to make use of big data. While the shortage of people with deep analytical talent is already well recognized (Davenport & Patil, 2012; McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012; Waller & Fawcett, 2013), data-savvy managers and analysts are now emerging as the missing link between analytical methods and a useful exploitation of the data in real applications. The education of data scientists is considered an important yet challenging task.
- Datacy: Giannotti et al. (2014) call education that improves general awareness of the importance and value of data, and gives a view of what data science is *datacy* in assonance with literacy. It describes the basic knowledge needed by citizens to comfortably live in a big data society.

Changes of the labour market due to the Internet have been discussed intensively. According to a study conducted by McKinsey Global Institute (2011), the Internet created 2.6 jobs for every 1 it eliminated. It is likely that it will be possible to observe a similar development in the context of the data economy.

Phenomena such as the increasing relevance of big data require a labour market to be fit in terms of structure and flexibility. While new jobs are created and new skills or skillsets become relevant, others become irrelevant.

4.3.5 Data market

4.3.5.1 Estimating the size of the data maket

A comprehensive market study on the European data economy addressing not only big data but also aspects of data reuse has not yet been published. Research on the data market, so far, has mainly focussed on open data in general and PSI in particular. Approaches to understand data reuse thus largely build upon work on the PSI market.

Accordingly, one of the first widely recognized studies focusing on economic aspects of the data market addressed the commercial exploitation of Europe's PSI (Pira International Ltd., University of East Anglia, & KnowledgeView Ltd., 2000). The authors of the study estimated a total value of PSI of €68 billion in the EU and €750 billion in the US in 1999. A more recent study estimated a market size of €48 billion for PSI in the EU (Dekkers, Polman, te Velde, & de Vries, 2006). This allows at least the conclusion that it is difficult to estimate the contribution made by PSI in particular and data reuse in general to the economy of the EU. Most other studies in the field of PSI estimate the value of a special field of PSI (e.g. geospatial, health) only (Vickery, 2011). It has to be considered, however, that the PSI market represents only a part of the data market.

In order to estimate the size of a market, it is necessary to know and understand the key actors in the market. According to Pira International Ltd. et al. (2000), a data value chain embraces actors dealing with data creation, aggregation, processing, and marketing/delivery as well as actors supporting them through providing the required infrastructure.

The overall data market and the market for PSI clearly differ in terms of who acts as a supplier of data. PSI is certainly only provided by the public sector. Moreover, the data markets are not so much organized in a linear fashion but rather like a network (de Vries & Hittmair, 2013) or in form of a lifecycle in which a value is continuously added to existing data (UK Data Archive, n.d.).

4.3.5.2 Building a business on data

It is not even close to be clear which business models will work in the data economy and which will not. In the end, it will be very much determined by the framework conditions given in certain country or group of counties. An end of trial and error in terms of business models is thus not to be expected soon. However, McCallum and Gleason (2013) published a description of the major categories of business models recently, which provides a useful overview of what will be possible in terms of business models in the data economy and what isn't.

- Collect/supply: Data is gathered by hand or through scraping[118] and sold to interested parties. As opposed to tangible goods, the same dataset can be sold over and over. While the costs of

[118] Scraping is a software technique for extracting information from websites and other sources.

gathering data may be high the marginal costs of distribution, if done electronically, are close to zero. The most relevant recurring expenses may be related to storage and bandwidth. The collect/supply business model is particularly relevant, if the data sold is a by-product of something else a company does.

- Store/host: While a traditional in-house system or self-managed cloud service makes sense for many companies, others may better offload the data management to third parties. Offloading data management may be particularly useful for data that is very large or otherwise difficult to store. In quite some industries companies are required to store data for regulatory purposes. The store/host business model gets particularly interesting if the same data should be hold for several clients. Offers can go beyond storing and providing access to raw data and also include host analysis services.
- Filter/refine: Problematic records or data fields are stripped out or interesting data subsets are released. Data with malformed, missing, duplicate or incorrect records is common. The basis of the filter/refine business model is to supply a clean dataset that removes or corrects rogue records. Processing bad records, however, is not trivial. It is critical to understand the problem domain and know the clients' interests.
- Enhance/enrich: The enhance/enrich business model is similar to the filer/refine business model. The idea is to spare clients the trouble of preprocessing data themselves. As opposed to the filter/refine business model, however, the strategy is not to subtract or normalise but to blend in other datasets to create a new picture. Public domain or other open datasets are ideal candidates for enhance/enrich business models as they generally have few restrictions on resuse. Enhance/enrich business models are in general particularly relevant from a data reuse perspective.
- Simplify access: Business models simplifying access to data help people cherry-pick the data they want in the format they prefer. The simplify access business model gets particularly interesting if several clients that need the same data wish to perform the same post-processing. Creating APIs such that clients can programmatically fetch the subsets they need in a machine-readable format may be an option in some cases. Business models of this type take advantage of the fact that collecting, segmenting, subsetting and cleaning data typically consumes a considerable share of the efforts related to a data-analysis exercise.
- Analyse: Focusing on analysis is another popular strategy for a business model in the data economy. Whereas the collect/supply business model is based on a straightforward transaction, data analysis is more of an indirect pursuit: Companies can offer to analyse someone else's data, they can analyse their own data and they can some data and sell the results. The unifying theme is to profit from insights that lie within a dataset. Generally, analyse business models are a broad and open-ended topic.
- Obscure: Companies wish to protect their data in-transit so that others do not profit from data exhaust or similar byproducts. Individuals, as they learn more about the ways in which

companies collect information about them, increasingly prefer to maintain their privacy. The focus of obscure business models is to inhibit people and companies from seeing or collecting certain data. Building tools to obscure information of otherwise foil data collections may be key offers of companies following the obscure business model.

- Consult/advice: The strategy behind consult/advice business models is to provide guidance on others' data efforts. A key element of a data consulting operation is domain knowledge beyond that of the client's industry. Consulting is certainly not unique to the data arena but it does have its own rules.

Although Americans seem to be somewhat more adventuresome in terms of new business models, the overall situation does not differ much between the US and Europe.

Concrete business models are also characterised by their cost and revenue structure. Business models can be cost-driven or value-driven and many fall in between the two extremes. Cost-driven business models aim to minimise all costs at every opportunity, for instance, by automating and outsourcing tasks. Value-driven business models focus on high-quality which allows higher prices for a desired level of service.[119] Fixed costs result mainly from hardware and software, office space, and salaries. Hardware and software costs can become variable to some extent if on demand services are used. Automation may to some extent lower the money spend on salaries.

The cost model of a company affects the price of the data sold. Companies try to take advantage of economies of scale and scope. Cost-driven companies may sell data at a price close to their marginal costs. Selling more data often reduces the proportional effort and thus allows lower prices. Value-driven business models may tend to be more complex with a higher amount of data. The effect of economies of scale are working better within the cost-driven business models. Both can benefit from economies of scope like the reuse of distribution channels.

4.3.5.3 Value and price of data

The value and price of pieces of data and entire datasets depends on various characteristics. Among the most relevant characteristics are:

- Quality of data (low proportion of malformed, missing, duplicate or incorrect records)
- Breadth of data
- Granularity of data
- Currency of data (up-to-date)
- Accessibility of data (several formats)

[119] http://www.mbaco.com/resources/finance/what-is-a-cost-structure

Data does have value because it improves decision making (Mayer-Schönberger & Cukier, 2013). Pricing is based on the question how much it costs for a potential client to gather the data by themselves. Other than that, the data market is determined by supply and demand. However, the fact that data is intangible and can be sold over and over affects this market mechanism.

4.3.5.4 Barriers and challenges

Key barriers with respect to the development of the data economy are the limited amount of high-quality data to build upon as well as problems related to data access.

Making data available for reuse works in different contexts but the motivations cannot be transferred easily to other contexts. Academic data sharing already works quite well, for instance, but it has to be understood that it is neither driven by monetary incentives not by the desire for scientific progress as a whole; it is driven by individual reputation (Fecher, Friesike, Hebing, Linek, & Sauermann, 2015).

The lack of data together with issues related to characteristics such as quality and currency lead to a pricing uncertainty. The update cycle of government data, for instance, is often very long. However, ensuring that it is regularly updated would most likely allow realizing increased uptake of the data (Capgemini Consulting, 2013). The lack of suitable hampers the development of data reuse and the data economy (Pira International Ltd. et al., 2000).

Consultative Committee for Space Data Systems (2012) and Hsu, Martin, McElroy, Litwin-Miller, and Kim (2015) postulate proposition on data reuse as well as the data economy in general and specify several requirements to be met:

- Data sharing/management policy pressure is needed to foster the opening of data
- IP has to be ensured for the data producers
- Licensing needs to be simplified
- Opening the complete metadata and the workflow documentation should be facilitated, for instance, through the development of standards
- Compensations for efforts need to go beyond reputation; data reuse has to be rewarded
- Cooperation between data creators and data users can be beneficial for both
- Data users need a possibility to assess the relevance of data
- Data users need to be able to determine whether the data were trustworthy (e.g. through certificates, open data documentation)
- Best practices in experimental methods and in the storage, archiving, and dissemination of experimental data should be applied
- There is a need for centralized data storage and marketplaces

- Standards or guidelines are needed to facilitate interoperability and reuse
- More frequent communication between investigators will lead to rescue of data and knowledge from inaccessible dark data storage and will facilitate exploitation and reuse
- Network activities would useful to support data reuse companies

4.4 Summary and conclusions

The section on socio-economic framework conditions first focuses on the societal perspective and deals with the economic perspective afterwards. Societal and economic aspects cannot always be separated clearly. The discussion on the societal perspective starts with the understanding of big data as a socio-technical phenomenon, continues with perceptions of big data issues and the psychology of mass self-communication. Subsequently, strategies for limiting big data analysis, the creation of a data culture and barriers of data incompleteness and the skills gap are addressed. Aspects of data incompleteness and the skills gap are also addressed within the scope of the discussion of the economic perspective. While the elaboration on the labour market, education and training is closely related to the skills gap, business models outlined within the scope of the description of the data market show how the issue of data incompleteness can be the core of companies active in the data economy. The discussion of the data market also focuses on aspects such as the size of the data market or the value and price of data. In addition to that aspects of financing economic activity related to data, the institutional landscape relevant for the data economy and the way from research and development to business sophistication are addressed within the scope of the discussion of the economic perspective.

5 Framework conditions: A technological perspective

5.1 Introduction

Within the last decade, technology has rapidly developed in terms of its significance in the European economy and is now a corner stone for all bigger companies in Europe. The decrease of costs for hardware and bandwidth together with the a broader availability of software and data management systems has allowed companies to collect and store data in a much more detailed and fine grained way than ever before. This opens a range of opportunities for Europe to use data in a meaningful way and to share and reuse it. As of today, European companies are under high pressure because of increasing competition with other global actors. In fact, Europe is clearly lacking behind in many aspects in the ICT industry: A range of big players such as Google, Microsoft, Apple or Yahoo are stemming from the US and China is quickly rising in importance with key players such as Baidu and Alibaba.

The possibility to collect, analyse, use and reuse data and the concept of building an economy around data may provide a key element for Europe as it allows the European society to benefit from one of its main competences: knowledge (European Union). In fact, data may be seen as “the new oil for the

digital age" as pointed out by Neelie Kroes, former Vice-President of the EC responsible for the Digital Agenda[120].

This section will give an overview of the current conditions that European actors are faced with in terms of the data economy from a technological viewpoint. This covers a range of topics reaching from describing the role of (de-facto) standards to describing security and privacy aspects.

5.2 Technology infrastructure

Fast access to information about customers and markets are a key enabler of success today. Technology is rapidly changing and influences the way business is made. For example, the iPad has been introduced only about 5 years ago. This introduction has significantly changed the way information is consumed today and has led to a whole range of tablets (Android tablets from different manufacturers, Surface from Microsoft, etc.) being available on the market today (Chen, Georg, & Loewi, 2011). This technological development affects both consumers and businesses that often use tablets to e.g. access machine information, to monitor production lines and to exchange status information with others.

Similarly to this example, technology is quickly evolving in all areas of business. This includes the availability of technical infrastructure in terms of bandwidth and hardware infrastructure. There are, however, big differences between different parts of Europe. While big cities often provide a high standard in terms of technical infrastructure, the countryside is often lacking behind. According to Stern, Adams, Alison E., and Boase (2011), rural communities are behind other types of places in terms of the availability and use of broadband high-speed technology. However, despite this gap, the technological infrastructure is generally increasing rapidly allowing European organizations to exchange even bigger amounts of information in a faster way than ten years ago. Data centres provide affordable infrastructure for companies that do not want on-site hosting of data and cloud providers such as Amazon allow companies to offload a broad range of their infrastructure to a hosting partner. Although each data center is a little different, the average cost per year to operate a large data center is usually between $10 million to $25 million. 42 percent: Hardware, software, disaster recovery arrangements, uninterrupted power supplies, and networking. 58 percent. Heating, air conditioning, property and sales taxes, and labor costs (Sekhar, Jeba, & Durga, 2012).

As a result, the technological infrastructure nowadays provides a solid ground for establishing a European data economy as it allows companies to actually acquire, manage and use data on a broad scale. This allows companies to develop new business models that go beyond the internal use of data by offering data to others and by consuming data as needed. Statista – an online statistics portal – is estimating this market to reach $50 billion in 2017, as shown in Figure 3[121].

[120] http://europa.eu/rapid/press-release_SPEECH-12-149_en.htm
[121] http://www.statista.com/statistics/254266/global-big-data-market-forecast/

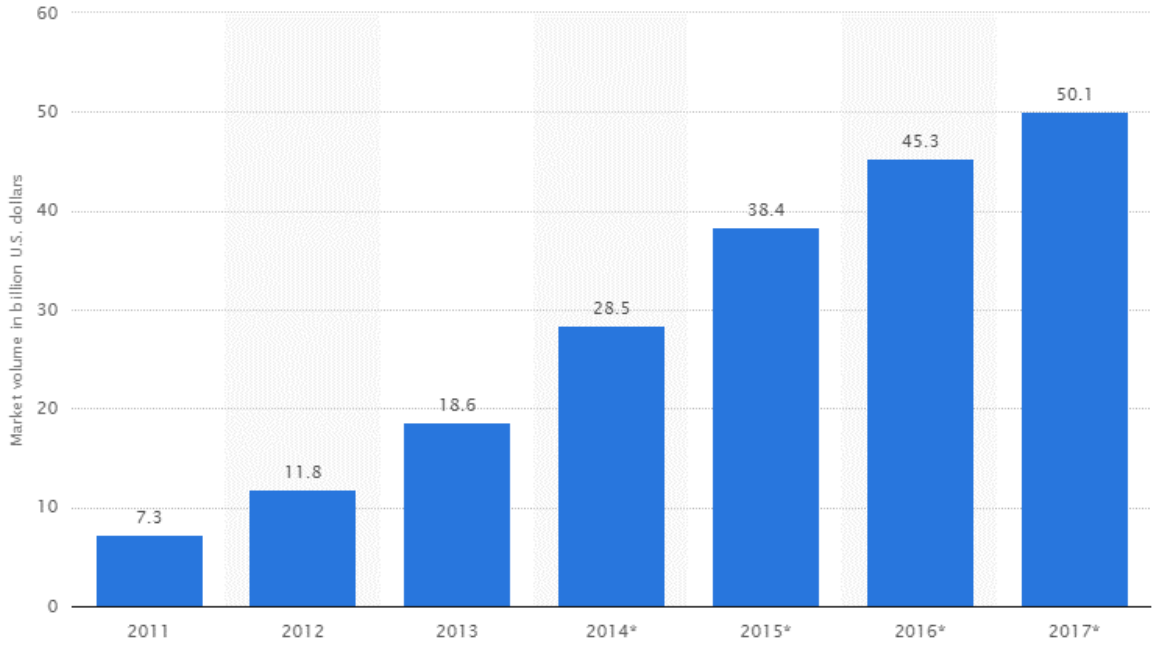

Figure 3: Big data market forecast worldwide from 2011 to 2017 (in billion US dollars)

The importance of cloud computing is increasing and it is receiving a growing attention in the scientific and the industrial community. The last three reports of the Gartner's List of Strategic Technologies (2013-2015)(Gartner, 2013, 2014, 2014) considered cloud computing one of the top 10 most important technologies and with solid prospect in successive years by companies and organizations.

5.3 Technical Characteristics of (big) data

The term *big data* may clearly be considered a buzzword today and a broad amount of projects and products can be found in this domain. The meaning of *big* obviously depends on the context and may be perceived differently for, for instance, a European SME and a big player such as Amazon or Google.

"Companies like Google, Facebook, Amazon, and eBay manipulate petabytes of data every day. For example, the retail chain WalMart handles more than one million transactions per hour, and manages databases with more than 2.5 petabytes of data. Facebook handles 20 petabytes of data, managing 20 billion photographs in 4 different resolutions, growing by 2 billion photographs per month. The Facebook database is serving 600,000 photographs per second for a user base of 300 million active users. Google manages vast amounts of semi-structured data: billions of URLs with associated internet content, crawl metadata, geographic objects (roads, satellite images, etc.), and hundreds of terabytes of satellite image data, with hundreds of millions of users and thousands of queries per second." (Feuerlicht, 2010)

Data does, however, not only have value to the European economy when it may be considered as *big* in terms of the amount of data. In fact, smaller datasets may deliver more value to their users in case the quality is higher. Within the EuDEco project, data and its use and reuse is therefore considered, independent of the size.

Figure 4 shows characteristics that have been adopted by the EU Big Data Value initiative[122] based on a characterization from Michael Walker[123].

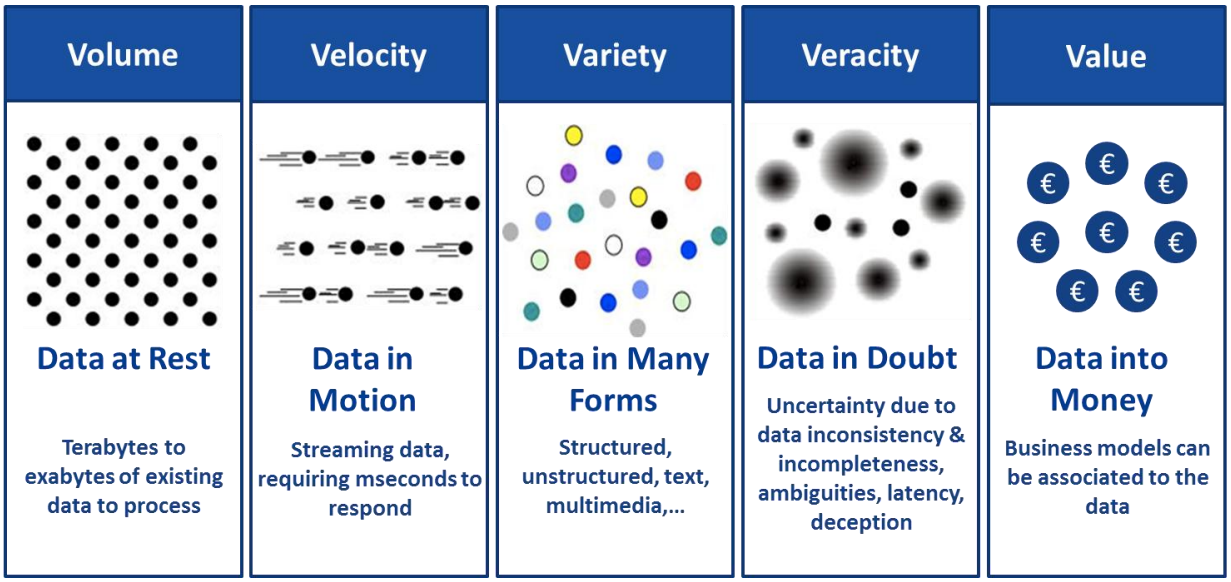

Figure 4: Characteristics of data[124]

5.3.1 (De-facto) standards and diversity

5.3.1.1 General perspective

The official motto of the EU is *united in diversity*. This motto reflects the broad diversity of Europe as a union containing many different nations with many different preconditions. In fact, this diversity of Europe with many different cultures, traditions and languages in Europe is seen by the EU as a positive asset for the continent[125].

This diversity is also an integrated part of the upraising data economy in Europe as different regions of Europe are marked by a whole range of different and sometimes overlapping *de jure* standards and *de*

[122] http://www.bdva.eu/
[123] http://www.rosebt.com/blog/data-veracity
[124] Source: BDVA slide set, S. Campbell for http://www.bdva.eu based on http://www.rosebt.com/blog/data-veracity
[125] http://europa.eu/about-eu/basic-information/symbols/motto/index_en.htm

facto standards. In many aspects of this domain, no European standard can be identified. Europe is embedded into a globalized world and as such, the technologies used within the area of big data are global as well. The complexity of modern technology, especially its system character and the globalization, have led to an increase in the number and variety of standards (Tassey, 1999). For example, many data formats exist which are nowadays used to exchange business information, ranging from electronic data interchange (EDI) to XML (Extensible Markup Language), JSON and even comma separated values (CSV). Nevertheless, standardization is an important aspect in ICT as it delivers a frame that companies can make use of such as XML as a widely accepted markup language standardized by the World Wide Web Consortium (W3C)[126]. From a technological level, a technology supported by a standardization body does not mean that it is globally used as the only solution but at least it means that this technology will have a wider acceptance and thus a chance of being used more often than others.

5.3.1.2 Official standardization organizations

From a technological perspective, many organizations define *de jure* standards (obligatory) or *de facto* standards (because of a dominant position). The following table shows a selection of prominent standardization organizations and their websites as a reference for further information:

CEN	
Full Name	Comité Européen de Normalisation
Nature	European standardization organization
Example Standards	DIN, EC (Eurocodes), CEN/TC*
Website	http://www.cen.eu

ECMA	
Full Name	Ecma International (formerly: European Computer Manufacturers Association)
Nature	Development and promotion of technical standards
Example Standards	ECMAScript, Office Open XML
Website	http://www.ecma-international.org

[126] http://www.w3.org/TR/2008/REC-xml-20081126/

IEEE	
Full Name	Institute of Electrical and Electronics Engineers
Nature	Association of engineers
Example Standards	POSIX, Bluetooth, Wireless LAN
Website	http://www.ieee.org

ISO & IEC	
Full Name	International Organization for Standardization International Electrotechnical Commission
Nature	International standardization development
Example Standards	ISO 9000, ISO 31000, ISO 3166, ISO 27001
Website	http://www.iso.org

OASIS	
Full Name	Organization for the Advancement of Structured Information Standards
Nature	e-Business and web service standardization
Example Standards	OpenDocument, DocBook, BPEL, ebXML
Website	https://www.oasis-open.org

W3C	
Full Name	World Wide Web Consortium
Nature	Development of standards for the world wide web
Example Standards	XML, HTML, SOAP
Website	http://www.w3.org

5.3.1.3 De facto standards

Within the ICT domain, many organizations are relying on *de facto* standard i.e. a format that has become a standard not because it has been approved by a standards organization but because it is widely used and recognized by the industry as being standard.. An example is the XLS format, the data format of Microsoft Excel, which is often used to exchange spreadsheet data, for instance, in the business world to share orders or delivery information.

Additionally, formats and standards are sometimes used for purposes different from those for which they have been developed. For example, PDF or CSV are commonly used formats in B2B (Business-to-Business) scenarios although they have not been developed and standardized for this purpose.

Many de facto standards have also been developed without the purpose to create a new standard in the first place. An example of a de facto standard for exchanging data between different systems is JSON. A format specification has now been initiated by two competing initiatives, one driven by Google [127] and one by ECMA[128].

5.3.2 Data collection and delivery

5.3.2.1 Description

Over the last years, the process of storing data got more and more inexpensive as storage technologies got more affordable. As such, many companies today store vast amounts of data. For example, data may stem from production processes – e.g. by adding sensors to the production line of an automotive manufacturer – or from eCommerce systems – e.g. by storing the customer behaviour. In general, before data is stored it needs to be retrieved, which may happen in two ways:

1. Data may be harvested by actively collecting data. An example for this approach is the active query of a sensor or the tracking of customer behaviour. Querying data may happen periodically (polling), in intervals (e.g. once per minute), based on events (e.g. when a sale is made) or via stream based data.
2. Alternatively, data may be delivered to a storage system, which may happen via service calls, for instance.

Retrieving data may also happen between two partners having a data supplier on the one hand and one or more data consumers on the other hand. As such, data may be created in one organization and may be consumed by another.

5.3.2.2 Challenges and barriers

The following table shows barriers and challenges that arise within this topic:

[127] http://tools.ietf.org/html/rfc7159
[128] http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

Challenge	Description
Stability of interfaces	Delivering or receiving data between partners requires high stability of their interfaces. This is challenging in a market where technology is constantly changing. As of today, APIs provided by big players constantly change, services are shut down and data formats are removed or extended. In 1:1 scenarios where two partners collaborate on the same level, interfaces are usually stable over a longer period of time and changes are communicated and discussed equally. However, in 1:n relationships, this is often not the case. As an example, big players such as Google or Amazon are constantly changing their APIs without the need to discuss those changes. In those situations, data consumers are forced to adopt their systems although long term backward compatibility is often not guaranteed.
Data format evolution	Similarly to changing interfaces, data formats may also change in terms of syntax and semantics. For example, the eCl@ss classification system[129] – a product classification and description standard for information exchange between customers and their suppliers – is widely used today and constantly changing so that partners exchanging eCl@ss data need to agree on a joint version of the standard.
Legacy systems	Collecting data is a case-specific exercise as data can have many diverse origins. It may be produced by a sensor, a webshop system or an Enterprise Resource Planning (ERP) system. In many cases, there is no standard available and most sensors have a sensor-specific protocol defined by the manufacturer, which then needs to be encapsulated manually in order to catch the data of the sensor. This is a time-consuming process, which needs to be repeated with each new data source.

5.3.2.3 Approaches and example cases

In order to overcome the challenges and barriers identified above, systems are often wrapped with an abstraction layer. From a technical perspective, this abstraction acts as a façade pattern (Gamma, Helm, Johnson, & Vlissides, 1994) hiding the underlying structure. The creation of an abstraction layer does, however, not prevent the re-creation of interfaces and integration work in case of changes but it ensures that the changes do not impact other data oriented systems as they are fully based on the abstraction layer.

[129] http://www.eclass.eu

Additionally, an active change management approach can help to reduce dependencies and to migrate to new interfaces if needed. Approaches such as the Duplo architecture pattern (Hasselbring et al., 2004) may help in terms of software migration.

Finally, a typical approach to reduce the impact of changing interfaces and data is the early announcement of changes combined with a parallel provision of old interfaces. Data and service providers such as Google tend to inform users of old interfaces about 6-12 months before switching them off[130]. This early *end-of-life*-notice helps partners to adopt their systems in time.

5.3.3 Scalability and data management

5.3.3.1 Description

Managing data and the information that comes along with it, is critical within a future data economy. The hardware got more and more inexpensive during the last decade and organizations are nowadays capable to store much more data for an affordable price. As such, scalability is an important factor in this domain. In this context, scalability refers to the ability to handle the growing demands for effectively storing and retrieving data. In a perfect scenario, scalability makes it possible to start a small system and expand it as needs grow (Grolinger, 2013). The capability to use a storage infrastructure that dynamically adapts to the needs is often referred to as elasticity (Trivedi, 2014). In a distributed data management system, data is usually partitioned to achieve scalability and replicated to achieve fault-tolerance (Agrawal, El Abbadi, Antony, & Das, 2010).

5.3.3.2 Challenges and barriers

The following table shows barriers and challenges that arise within this topic:

Challenge	Description
Many data types	The term *data* covers many different types of data. This includes structured data (e.g. table-based data, JSON data), unstructured data (e.g. PDF or WebM files) and even semantic data (e.g. OWL (Web Ontology Language) based). These different data types obviously need different ways for storing and accessing them. As such, there is no universal *one fits all* solution that participants in a European data economy could make use of. Instead a case-by-case analysis is required for finding the right storage for each situation.

[130] http://googlegeodevelopers.blogspot.de/2014/12/announcing-deprecation-of-google-earth.html

Unknown needs	Unlike many other domains, the IT industry allows market actors to quickly enter the market and to offer new products in a very fast way – often without the need to invest into sophisticated machines or infrastructure. Many systems may be offered on the web and only require small web servers starting at less than €50 per month. However, as web systems get more and more popular, they often face a scalability problem in terms of storing the increasing amount of information and also in terms of offering their services. A good example is Wunderlist, which made its story public[131].
Cost prediction	Analogous to the previous challenge, costs are often unpredictable. In many cases, data storage scalability problems can be outsourced to business partners – e.g. by using a cloud storage space such as Amazon S3[132].
Conflicting data	One façade of a European data economy is that data may be exchanged in different ways. For example, data that is exchanged between company A and B may be copied, linked from A to B or synced (unidirectional or bidirectional) between the partners. Data that is copied may lead to outdated or even conflicting data.

5.3.3.3 Approaches and example cases

There is no *one-fits-all* solution for managing data in general as described above. Clear guidelines could be valuable for companies to select the right approach for a specific situation. This would allow companies to get an overview of the market solutions and to choose the right technology for their needs.

In many cases, it is not necessary to build a new infrastructure as cloud providers may offer good and mature data management solutions. Amazon is a leading provider of cloud services. The company offers a simple storage service called Amazon S3 that provides web service interfaces for REST (Representational State Transfer) and SOAP (Simple Object Access Protocol) that can be used to store and retrieve any amount of data, at any time, from anywhere on the Web.

However, in some cases, outsourcing to cloud providers may not be an option – either because of costs, political issues, data security and privacy aspects or technological aspects. In those cases, many popular solutions may be used which have been established in the last years.

For example, many big company executives see Hadoop as a low-cost alternative for the archival and quick retrieval of large amounts of data. Apache Hardoop is an Apache Foundation software being used by several large companies and institutions (Facebook, Yahoo, Linkedin, etc) (Donvito, Marzulli, & Diacono, 2014). Apache Hadoop provides reliable, scalable as well as distributed computing, and its library is a framework which allows the distributed processing of large data sets across clusters of computers using simple programming models. It provides a reliable shared storage with HDFS (Hadoop

[131] https://www.wunderlist.com/blog/improving-wunderlists-sync-part-2/
[132] http://aws.amazon.com/s3/

Distributed File System) and analysis with MapReduce (a distributed data processing model) system for large-scale data processing (Kumar, Parashar, Gupta, Sharma, & Gupta, 2014). Accordingly to Khan et al. (2014), one of the benefits of Hadoop is the capacity of processing extremely large volumes of data with varying structures (or no structure at all). Figure 5 shows an example of a data warehouse process chain using Hadoop.

Figure 5 An example of a data warehouse process chain using Hadoop (Davenport & Dyché, 2013)

As a second example, GlusterFS can be named as a general purpose distributed file system for scalable storage. It aggregates various storage bricks into one large parallel network file system (Allan, 2.010). GlusterFS is based on a stackable user space design for storing and managing binary files. GlusterFS can scale up to several petabytes on commodity hardware by avoiding bottlenecks that normally affect more tightly coupled distributed file systems.

5.3.4 Data transmission and exchange formats

The literature on data transmission is very broad. The literature focuses mostly on telecommunications and electrical engineering aspects which correspond to the physical layer (e.g., channel coding, multiplexing) and the data link layer (e.g., error detection, flow control) of the ISO/OSI reference model. Several relevant aspects for building a data economy model are covered under the umbrella term *computer networking*, which include communication protocols, routing, network performance (e.g., Guo et al., 2009; Wang & Ng, 2010) and security (see section 5.3.8).

Data exchange formats define the characteristics of data exchanged between parties. Most literature on data exchange formats focuses on the description or application of XML-based standards for data exchange, which are often tailored to very specific needs. Specifications were applied in different areas such as bioinformatics (Kalas et al., 2010), engineering projects (Barth & Fay, 2010) and graphs (Holt, Schürr, Sim, & Winter, 2006). Moreover, several publications are focused on XML itself or alternatives used for data exchange such as JSON and CSV. Finally, a few surveys have been documented which are

explicitly focused on the data exchange formats used by data marketplaces (e.g., Graves & Bustos-Jiménez, 2014; Moyle, Blackall, Hudson, & Mason, 2012; Schomm, Stahl, & Vossen, 2013).

Data transmission is of prime importance in a data economy that is characterised by a large number of actors interacting with each other. The use or reuse of data, particularly if more than one actor is involved, almost always requires some kind of data transmission, and each and every data transmission requires an agreement on a data exchange format.

5.3.4.1 Description

Data is available in numerous formats. According to a survey conducted by Schomm et al. (2013), data can be obtained from marketplaces in formats such as XML, CSV/XLS, JSON or RDF (Resource Description Framework). In addition to that, marketplaces also make data available in form of reports. If data is obtained in report form, customers usually do not have insight into the underlying raw data. Visual reports may be made available as PDF or XLS(x) files (MS Excel spreadsheets).

- XML: XML is a widely established standard for data transfer and representation that aims at being both human and machine-readable. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures. The characters making up an XML document are divided into markup and content, which are distinguished by the application of simple syntactic rules. XML is closely described by the W3C (Bray et al., 2006).
- CSV/XLS(X): As most structured data are laid out in a tabular way, it makes sense to wrap it into a table file format. CSV files store tabular data in plain text where each line of the file is a data record and each record consists of fields that are separated by commas. XLS is the binary file format used by MS Excel up until the 2007 version of the software. Later version of MS Excel use the file format XLSX which is based on Office Open XML, a zipped, XLM-based file format.
- JSON: When using JSON, data are presented as text in key-value pairs. By means of RFC (Request for Comments) 7159, the Internet Engineering Task Force (IETF) has recently published a comprehensive specification of the JSON format (Bray, 2014). Severance (2012) discusses the advantages of XML and JSON for data applications. According to him, industry support for XML is much broader than for JSON. JSON, however, represents common programming structures in a more natural and direct way than XML. Both XML and JSON allow the specification of schemas for validating the structure of data.
- RDF: The RDF uses subject-predicate-object triples to make statements about resources. RDF concepts as well as the syntax have been specified by the W3C (Klyne, Carroll, & McBride, 2014). There are several common serialization formats for RDF including Turtle, N-Triples, JSON-LD. N3 and RDF/XML. Although RDF/XML was the first W3C standard RDF serialisation format, other RDF serialisations are now preferred by many users (Miličić, 2011).

Among the data marketplaces investigated by Schomm et al. (2013) CSV/XLS(X) was the most widely used output format. However, most marketplaces offer data also in other formats such as JSON or XML.

The most popular way of data access is via an application programming interface (API). Alternatives to APIs are:

- the download of files,
- the use of specialized software and
- Web interfaces that display the data.

An API usually produces XML or JSON output. Schomm et al. (2013) consider that offering many ways to access data is a key feature of a data marketplace as it broadens the range of possible users.

Graves and Bustos-Jiménez (2014) have investigated what data formats are most often used to publish open government data. They looked at Data.gov[133] and Data.gov.uk[134], two of the largest government data portals, and found out that most data is published in XML, CSV, PDF and JSON. In many cases, data were published in multiple formats. Graves and Bustos-Jiménez (2014) also mention HTML (Hypertext Markup Language) and ZIP as formats but acknowledge that both are not data formats *per se*. While HTML means that a Web interface is provided or that data is made available directly on a website, ZIP files are actually archives containing other files.

Moyle et al. (2012) have also analysed the formats used for publishing open data. In addition to the formats already mentioned, they consider KML/KMZ and TXT particularly relevant.

- Keyhole Markup Language (KML) is an XML-based standard for expressing geographic annotation and visualization that is used, for instance, in Google Earth and Google Maps. It was created by a company which was later acquired by Google. Google Earth was the first program able to view and graphically edit KML files. Nowadays, KML is supported by other applications too. KMZ files are zipped and consist of one KML file as well as supporting files.
- TXT files are raw text files.

The total number of XML-based standards is very high. Most of them focus on quite specific needs. KML is only one example.

From a data transmission perspective, key aspects are usually somehow related to performance, security and cost. Performance gets even more important when the amount of data transmitted gets bigger as well as when data processing gets increasingly time-critical. More and more data applications meet both criteria and both are part of most definitions of big data. In terms of performance, throughput is what one would look at first. Throughput, however, is not just a question of physical link speed. Factors such as latency, error rates and window size affect throughput. Different protocols such

[133] http://www.data.gov/
[134] http://data.gov.uk/

as TCP (Transmission Control Protocol) and UDP (User Datagram Protocol) affect throughput based on these factors (Fujitsu, 2013).

From a data exchange perspective, with respect to the file formats discussed above both syntax and semantics are relevant. To ensure that data can be reused, it is essential that a sender can bring the data in a form that is understood by the receiver. Syntax describes the rules that have to be met when creating valid files of a given data exchange or file format. Semantics relates to the meaning of data that is in a syntactically valid format. Enhancing data with additional data is a way to allow receivers to make sense of data more easily. The semantic web community has been trying to make web pages understandable for computers by adding metadata (Berners-Lee, Hendler, & Lassila, 2001). RDF has been used a lot in the context of the semantic web idea.

5.3.4.2 Challenges and barriers

The following table shows barriers and challenges that arise within this topic:

Challenge	Description
Low throughput	Transfer of large amounts of data is a throughput bottleneck, particularly in the context of big data and real-time data applications. This is a key issue to be kept in mind with respect to the use of cloud services and other forms of locally distributed data processing and storage.
Large number of formats	Although it seems that there are rather few central formats, this impression is deceptive. Even if, for instance, XML or JSON files are used to exchange data, there are still many different standards for one and the same application. Everybody is free to develop XML and JSON schema documents defining what is syntactically right or wrong. Agreeing on standards is thus important in the context of data reuse.
Difficult processing	Some formats are difficult to process: if, for instance, sophisticated MS Excel spreadsheets or PDF files are used, the raw data does not become accessible easily to receivers. While data in report format can usually be well read and used by people, it is not of much use for computers. Reusing such data in an automated fashion is very difficult.
Difficult representation	Not all data can be easily represented in the main formats introduced above. The representation of video, image or audio data, for instance, is difficult.

5.3.4.3 Approaches and example cases

Measures are taken to continuously improve throughput. Recent literature mostly focuses on wireless networks (e.g., Katti, Katabi, Hu, Rahul, & Médard, 2006). Nevertheless, the transfer of large amounts of

data remains a bottleneck. Consequently, it is important that within the scope of data applications as much data as possible is processed right where it is collected and stored, and from where the results are taken.

The large number of formats and difficulties with respect to the processing of some formats will most likely remain challenges for the foreseeable future.

Moreover, there are data exchange tools and file format conversion utilities (e.g., Kolaitis, 2005).

5.3.5 Data aggregation, analysis and interpretation

Data aggregation, data analysis and data interpretation focus on making sense of data.

- Data aggregation describes the gathering of data as well as its expression in summary form. Online analytic processing is a simple type of data aggregation. Aggregated data can be the starting point for further analysis.
- Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different domains. The goal of data analysis is to discover useful information. Data mining is a common data analysis technique.
- Data interpretation has much to do with the discovery of useful information. To discover useful information, data has to be interpreted in the light its origin, processing steps carried out and the given context.

The terms are closely related and a clear distinction between the underlying concepts is hardly possible. With respect to the analysis of literature focusing on the topic, a distinction between the terms does not seem reasonable. The discovery and communication of meaningful patterns in data is under permanent development. Recent developments in the context of data analysis can be considered key drivers of many business models relevant in the context of the data economy. There is plenty of literature discussing data analysis aspects.

As data analysis is required to make sense of data, the success of the data economy heavily depends on the ability to analyse data effectively. Innovative data analysis approaches and algorithms are often critical for the success of data reuse in new contexts.

5.3.5.1 Description

Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories. There are several phases that can be distinguished. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases. Statistician John Tukey defined data analysis in the early 1960s as, “Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or

more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." (Tukey, 1962)

The tools available to handle the volume, velocity, and variety of big data have improved greatly in recent years. In general, these tools are not prohibitively expensive, and many of them are open source. Hadoop, the most commonly used framework, combines commodity hardware with open-source software. It takes incoming streams of data and distributes them onto cheap disks; it also provides tools for analyzing the data. However, these technologies do require a skill set that is new to most IT departments, which will need to work hard to integrate all the relevant internal and external sources of data. Although attention to technology isn't sufficient, it is always a necessary component of a big data strategy (McAfee et al., 2012).

There are three key types of analytics to be distinguished (Bertolucci, 2013)

- The purpose of descriptive analytics is to summarize what happened – like simple event counters. Is has been estimated that around 80% of business analytics is descriptive.
- Predictive analytics utilises a variety of statistical, modelling, data mining and machine learning techniques to study recent and historical data in order to allow making predications. Sentiment analysis is a common type of predictive analytics.
- Prescriptive analytics goes beyond descriptive and predictive models by recommending one or more courses of action and showing the likely outcome of each decision. Prescriptive analytics is a type of predictive analytics. Prescriptive analytics requires a predictive model.

Gandomi and Haider (2015) highlight the point that predictive analytics, which deals mostly with structured data, overshadows other forms of analytics applied to unstructured data, which constitutes 95% of the data considered big data. Gandomi and Haider (2015) reviewed analytics techniques for text, audio, video, and social media data, as well as predictive analytics.

The volume of data operated upon by modern applications is growing at a tremendous rate, posing intriguing challenges for parallel and distributed computing platforms (Kambatla, Kollias, Kumar, & Grama, 2014). In the past, RDMS were used to store and analyse data. RDMS however faced several problems as the volumes of data increased, analysis results became more time critical and data heterogeneity grew. Hadoop provides a reliable, shared storage and analysis system for large-scale data processing, where storage is provided by HDFS and analysis by MapReduce. While RDMS are good for real-time data retrieval with low-latency and small datasets, MapReduce is better capable to analyse larger datasets in a batch fashion. Hadoop became a top-level project of the Apache Foundation and is now used by hundreds of organisations (Savvas, 2013). The Hadoop components HDFS and MapReduce are complemented by Hive (a distributed data warehouse providing an SQL-based query language), HBase (a distributed column-based database) and Pig (a data flow language and execution

environment). There is plenty of literature discussing Hadoop and its components in detail (e.g., Dean & Ghemawat, 2010; Shvachko, Kuang, Radia, & Chansler, 2010; White, 2012).

5.3.5.2 Challenges and barriers

The following table shows barriers and challenges that arise with respect to analytic processes:

Challenge	Description
Scaling	A critical issue is whether or not an analytic process scales as the data set increases by orders of magnitude. Every algorithm has a point at which its performance ceases to increase linearly with increasing computational resources and starts to plateau or peak, turn over and start decreasing (Kaisler, Armour, Espinosa, & Money, 2013).
Noise accumulation	Analyzing big data requires the simultaneous estimation or test of many parameters. Estimation errors accumulate when a decision or prediction rule depends on a large number of such parameters. Such a noise accumulation effect is especially severe in high dimensions and may even dominate the true signals (Fan, Han, & Liu, 2014).
Spurious correlation	Spurious correlation, the fact that many uncorrelated random variables may have high sample correlations, may cause false scientific discoveries and wrong statistical inferences (Fan et al., 2014).
Incidental endogeneity	In a regression setting, incidental endogeneity means that some predictors correlate with the residual noise. The exogenous assumption that the residual noise is uncorrelated with all the predictors is crucial for validity of most existing statistical procedures. This assumption is easy to be violated as some of variables are incidentally correlated with the residual noise, making many procedures statistically invalid (Fan et al., 2014).
Heterogeneity	The main advantage brought by big data is to understand the heterogeneity of subpopulations which is not possible when sample size is small or moderate. However, big data requires sophisticated statistical and computational methods. The estimation procedure needs to be carefully regularised to avoid overfitting or noise accumulation, and to devise good computation algorithms (Fan et al., 2014).

The following table shows barriers and challenges, identified with the scope of unpublished expert interviews, related to big data solutions:

Challenge	Description
Sophistication	Big data solutions available today do not support concepts required to carry out specific clustering, classification or network analysis tasks. If one goes beyond simple aggregations and transformations of data, if use data mining methods should be used which are required, for instance, for carrying out predictive or prescriptive analytics, the limits of current big data technologies are reached quickly. Hadoop, for instance, does not natively support iterations and, consequently, does not allow executing iterative algorithms efficiently.
Usability	Big data solutions available today cannot be used by most of the *normal* staff in the organizational departments. Although there is a plethora of different solutions in the context of big data, hardly any solution makes its users exempt from the need to deal with system programming. System programming, however, requires skills and expertise hardly available in organizations. Additionally, most technology providers introduced their own languages. The lack of a common language did not only hamper the emergence of querying and visualization tools in the past but also further increased the requirements imposed on data scientists.

5.3.5.3 Approaches and example cases

Solving the problem related to analytic processes requires the rewriting of algorithms or the development of entirely new algorithms. Progress has been impressive over the past years but there is still potential for improvement.

Also with respect to big data solutions, issues such as the support of sophisticated analyses or the usability for normal staff have been improved. However, there is still much to do.

5.3.6 Reliability and availability

5.3.6.1 Description

Agrawal et al. (2010) considers reliability a key requirement to ensure continuous access to a service and defines reliability as the probability that a given application or system will be functioning when needed as measured over a given period of time. The same author also mentions that reliability is the capacity of ensuring safety and persistence of data in the presence of different types of failures.

Grolinger (2013) reported that another important factor regarding scalability is replication: storing the same data on multiple servers so that read and write operations can be distributed over them. Replication also has an important role in providing fault tolerance because data availability can withstand the failure of one or more servers. Furthermore, the choice of replication model is also

strongly related to the consistency level provided by the data store. For example, the master-slave asynchronous replication model cannot provide consistent read requests from slaves. Master/slave is a model of communication where one device or process has unidirectional control over one or more other devices or processes.

In this aspect, the terms *reliability* and *availability* are often seen as related and are often mixed. In the course of EuDEco, the terms can be defined as follows

- Reliability: The reliability of a system is the conditional probability that the system survives for a determined interval, given that the system is operational at a specific period (Siewiorek & Swarz, 2014).
- Availability: The availability of a system is the probability that the system is operational at the instant of time. The rate of failure/maintenance events and the speed of recovery are relevant criteria (Griffith, Virmani, & Kaiser, 2007; Siewiorek & Swarz, 2014)..

The Internet itself is constructed as a set of distributed servers with data being physically stored at very different locations. Systems relying on data and data consumption often follow a similar approach by distributing data to several locations – either by storing different parts of data on different locations or by replicating data in order to achieve fault tolerance towards data storage.

5.3.6.2 Challenges and barriers

The following table shows barriers and challenges that arise within this topic:

Challenge	Description
Data loss	Data that is not replicated or synchronized with another data source may be lost in case of a hardware or data failure. Most cloud storage providers therefore use data replication techniques to avoid this. Most servers provide RAID (Redundant Array of Independent Disks) systems to lower the effect of hardware failures. Nevertheless, data loss gets especially important and challenging in terms of big amounts of data and especially in self-hosted and distributed environments involving many nodes.
Split brains	Data replication and distribution often provide a good way to avoid data loss. They do, however, introduce new challenges from a technological perspective. For example, a replicated data storage among two servers may face a so called *split brain* situation in case that both loose the connection to each other while still being online and available to others. In that case, their data may become inconsistent and an automatic merging may not be possible anymore. To avoid this, most replication scenarios require the involvement of more than two different replication destinations.

Chain reactions	Within a European data economy, a promising approach could be to combine data from different data providers without hosting data at own servers. In that case, information systems may consist of a combination of multiple providers – e.g. by integrating data from Google, Facebook, Twitter and others. However, this live-coupling also leads to an increased dependency on the systems: If one external system becomes unavailable, the full system may become unavailable as well. Thus the chances of unavailability increase with the number of depending external systems for data provision.
Failover	To avoid a low availability of systems and their data, failover approaches are needed which can instantly replace a damaged system/storage without long down times. However, corresponding systems often do require detailed technological expertise.

5.3.6.3 Approaches and example cases

Within the last years, a set of technologies has been established to face the challenges that appear with the increased amount of data and the tendency of having multi-node server environments. A few examples shall be named in this overview:

In the area of database storage, SQL (Structured Query Language) is often used to manage information. A popular example to increase the reliability and availability of SQL databases is Galera[135] – a replication solution which can extend popular databases including MySQL and MariaDB. It provides master-master replication running on traditional servers and allows fast and ad-hoc replication.

For file storage, GlusterFS[136] may be named as outlined in section 5.3.3. Ceph[137] follows a similar approach by providing a storage system called RADOS, which provides file storage as well as an Amazon S3 compliant interface for structured and unstructured data.

As a third candidate, MongoDB[138], provides a distributed NoSQL approach which does not only provide replication (as replica sets) but also horizontal sharding (a database is broken down into smaller chunks called *shards* and those are spread across a number of distributed servers) for distributing the work load and increasing the scalability and finally the availability of a data storage oriented system. It includes an automatic failover strategy based on master-slave replication.

[135] http://www.galeracluster.com
[136] http://www.gluster.org/
[137] http://ceph.com/
[138] https://www.mongodb.org/

5.3.7 Technological responsibility

Responsibility for technology is transferred whenever third-parties are involved in technology-related business processes. Outsourcing plays an important role in the context of IT. In principle, outsourcing is a way of seeking for help. Particularly in emerging fields such as insight extraction from data, where experts are rare and expensive, it is often the case that third parties are more capable or familiar with certain practices. Responsibility is a key issue when data-related activities are outsourced to third-parties.

In the context of the data economy, using cloud infrastructure or software services became very popular over the past couple of years (Hashem et al., 2015). Using them has made business models worth thinking about that were considered not economically viable or too risky before. Cloud services are as relevant for data reuse as they are for data use. To succeed, in both cases the actors involved have to address issues related to technological responsibility.

The variety of literature on IT outsourcing and cloud services in the context of big data is quite broad. Apart from scientific articles, there are numerous publications from service providers and technology manufacturers addressing aspects related to their products and services as well as more general challenges.

5.3.7.1 Description

Cloud-based solutions combine different concepts and technologies that have been discussed in the fields of big data and cloud computing. Many popular solutions which support the analysis of data have been made available as cloud services in the meantime. Solutions are not only being made available by long-established providers of database and data warehouse solutions but also by innovative startups.

The variety of solutions related to big data is broad. There are comprehensive solutions that support all major activities from data access and integration via storage and analysis to visualisation as well as highly specialised products that support only a few activities or sub-activities. Big data products are usually designed in a way so that they can be integrated with other products or solutions easily. Solutions and products are usually provided as software only, as a combination of hardware and specifically designed software, and as cloud services. With respect to public cloud services relevant for big data, there are IaaS (Infrastructure as a Service) and SaaS (Software as a Service) offers. Software which can be operated on private cloud environments is not considered here as it is not relevant in the context of technological responsibility. While IaaS offers include infrastructure, a Hadoop cluster, for instance, only, SaaS offers include full-fledged big data solutions or products.

Among the most popular IaaS offers are:

- Amazon Elastic MapReduce (EMR)[139] provides Hadoop as a cloud service. As the cluster used consists of Amazon Elastic Compute Cloud (EC2) instances it is scalable. EC2 is a computation service provided by Amazon. Installing tools such as Hive, Pig or HBase is possible. Third-party vendors provide graphical user interfaces that can be connected to EMR. Additionally, Amazon offers various storage, database and data transmission services.
- Google BigQuery[140] is a cloud service that allows queries on large amounts of data using a SQL-like language. The data to be queried must be available through Goole's Cloud Storage. Various third-party vendors provide graphical user interfaces and tools that can be connected to BigQuery.
- Microsoft HDinsight[141] provides Hadoop on the basis of Azure as a cloud service. Azure is Microsoft's cloud platform. Business intelligence tools from Microsoft such as PowerPivot and Power View can be used to analyse data.

Examples for popular SaaS offers are:

- MicroStrategy Cloud[142] is a business intelligence offer that is provided as a cloud service. It can be connected to local data sources as well as to cloud sources. It is also possible to use storage services provided by MicroStrategy. MicroStrategy Cloud uses in-memory technology.
- Tibco Spotfire Cloud[143] is a business intelligence solution from Tibco that is provided as cloud service.
- With Forward Demand[144] Blue Yonder offers a demand forecast solution taking sales, price, promotion and weather data into account. The solution can be connected to ERP and data warehouse solutions via interfaced.

Apart from business intelligence tools that can be connected with big data solutions, there are also solutions directly aiming at big data. They are referred to as big data suits and are usually based on Hadoop. Such solutions usually combine different big data tools and cover everything from data access and integration via data storage and analysis to data visualisation. Graphical user interfaces usually allow the visual mapping of big data sources and targets.

[139] http://aws.amazon.com/elasticmapreduce/?nc2=h_ls
[140] https://cloud.google.com/bigquery/
[141] http://azure.microsoft.com/en-us/services/hdinsight/
[142] http://www.microstrategy.com/us/platforms/analytics/cloud
[143] http://spotfire.tibco.com/discover-spotfire/spotfire-overview/spotfire-cloud
[144] http://www.blue-yonder.com/en/products/forward-demand.html

Examples for popular big data suits are:

- The analysis platform Datameer[145]allows the performance of large-scale data analyses. The platform is based on Hadoop and the user can choose the distributions to be used. Even the use of the cloud service EMR is possible. Both structured and unstructured data can be processed.
- Talend[146] also combines various big data tools in one analysis platform. With more than 800 connectors Talend can be connected to a huge number of data sources including, for instance, Google BigQuery. Talend support different Hadoop distributions.

5.3.7.2 Challenges and barriers

IT outsourcing, which implies shifting technological responsibility to a third party, promises benefits such as increased cost predictability and reduced costs, higher flexibility and scalability upon demand. Organizations trying to realize these benefits, however, face several challenges and barriers. These challenges and barriers affect organizational performance adversely in case they are not properly addressed.

Armbrust et al. (2010) compiled a list of the top 10 obstacles to and opportunities for growth of cloud computing. Some of the items are highly relevant in the context of technological responsibility and thus mentioned here. The following table shows general barriers and challenges that arise within this topic:

Challenge	Description
Availability	With business continuity in mind, service availability is a major concern as soon as responsibility for critical processes is transferred to a third party. Users expect high availability from services which is difficult to guarantee. Technical issues of availability aside, cloud providers may go out of business or are the target of regulatory action (Armbrust et al., 2010). The use of service-level agreements is common in the context of availability. Problems related to availability have to be taken into account when defining a business continuity strategy.

[145] http://www.datameer.com/
[146] http://www.talend.com/products/big-data

Vendor lock-in	Vendor lock-in makes a customer dependent on a vendor for products and services, unable to use another vendor without substantial switching costs. A lock-in situation is also possible with respect to certain technologies. Software stacks have improved interoperability among platforms but most APIs for cloud services are still proprietary. Customers face difficulties when trying to extract their data from one site to run on another. Lock-in may be attractive to providers of cloud services but it makes users more vulnerable to price increases and availability issues.
Performance unpredictability	The performance of the infrastructure is particularly important if applications are time critical. Infrastructure performance is discussed in sections 5.2 (general) and 5.3.4 (with a focus on computer networks). According to Armbrust et al. (2010), multiple virtual machines can share central processing units (CPUs) and main memory surprisingly well in cloud computing, but that network and disk sharing is more problematic.
Security and compliance	Security and regulatory compliance are key issues if technological responsibility is transferred to third parties. They are among the most often-cited objections to cloud computing (Armbrust et al., 2010). Therefore, particular attention is placed on this aspect (see table below).

The following table shows challenges and barriers identified by Thalmann, Bachlechner, Demetz, and Maier (2012). The challenges and barriers are particularly relevant from a security and compliance perspective.

Challenge	Description
Auditing cloud environments	A deep understanding of clients' IT including all outsourced IT services and access to audit evidence are prerequisites for auditors to conduct meaningful examinations. However, restricted access to and control of outsourced IT services makes it difficult to develop the confidence required to ascertain that requirements are met for the organization's IT infrastructure as a whole. This is particularly true if IT services are outsourced to service vendors which subcontract other vendors and if public clouds are used.
Managing heterogeneity	Effectively integrating IT services provided by multiple service vendors is difficult due to the considerable heterogeneity among IT services. Currently, there are hardly any standardized interfaces for IT services available. The inability to integrate IT services gives rise to problems.

Coordinating involved parties	It is difficult for service vendors to adapt to their clients' ever-changing requirements, particularly when taking into account the IT services they outsource to other vendors. As a consequence of differing national legislations, it is not uncommon to see conflicting client requirements in complex IT outsourcing arrangements.
Managing relationships	Contractual agreements such as service-level agreements (SLAs) are a primary means of documenting the agreements reached by the different organizations involved in IT outsourcing arrangements. Monitoring compliance with such agreements, however, is difficult for client organizations, particularly, if their vendors in turn subcontract other service vendors.
Localizing and migrating data	The place where data is stored or processed is important for organizations involved in IT outsourcing arrangements, particularly where there are differences in national legal obligations. Ownership of data is another issue that needs particular attention in the context of IT outsourcing. Keeping track of data is particularly difficult for clients if vendors subcontract other service vendors.
Coping with lack of security awareness	As in many other contexts, people play a major role in IT outsourcing arrangements. In addition to contractual agreements and technical measures, service vendors also need to ensure that their workforce takes security and compliance seriously. Achieving and maintaining awareness proved to be difficult enough within organizations but is exacerbated in complex IT outsourcing arrangements.

5.3.7.3 Approaches and example cases

The only way to ensure very high availability of cloud service is to use multiple providers. The high-availability computing community has long followed the idea of avoiding single points of failure. A solution to vendor lock-in would be to standardise APIs. Armbrust et al. (2010) proposes to offer something like *gang scheduling* to deal with performance unpredictability in cloud computing. Gang scheduling is a scheduling algorithm for parallel systems that schedules related threads or processes to run simultaneously on different processors.

Apache jclouds[147] is an open source multi-cloud toolkit that allows creating applications that are portable across clouds while giving full control to use cloud-specific features. The toolkit is supported by

[147] https://jclouds.apache.org/

about 30 cloud providers and cloud software stacks including Amazon, Azure, and Googe, and provides several API abstractions as Java and Clojure[148] libraries.

CloudRAID[149] aims at mitigating risks that come with cloud computing, particularly for storing large amounts of data, such as unauthorized access to storage resources or dependence on specific service providers. A RAID algorithm is applied on encrypted files in order to calculate the parity chunks, which are separated from the data chunks. The resulting chunks are then distributed to different cloud storage repositories.

5.3.8 Security and privacy aspects

5.3.8.1 Description

A growing number of massive data breaches are degrading the personal privacy of people around the world. Consequently, data security and privacy policy are ongoing concerns in Europe (Howard & Gulyas, 2014). In today's information age, data is an important asset of each organization. The security of the data is also vital in the industry (Ravikumar, Manjunath, Ravindra, & Umesh, 2011).

"The large usage of the Internet and social networking in particular brings with it the need to design and implement a large set of security enhancing and privacy preserving protocols and standards. Several protocols and security mechanisms have been proposed to ensure primary security features such as confidentiality, integrity, authenticity and non-repudiation. However, ensuring the correctness of these protocols is crucial in ensuring user confidence in system security." (Pai, Sharma, Kumar, Pai, & Singh, 2011)

The prevention of security issues mostly involves identifying unique threats and challenges which need to be addressed by implementing the appropriate countermeasures. This is especially important within a future European data economy as such an economy will involve multiple actors and thus multiple security violation possibilities. Information security, including different aspects such as integrity, confidentially, privacy, and assurance, is a major concern for all countries, including the developed ones (Tongia, Subrahmanian, & Arunachalam, 2005).

In a public cloud scenario, data is often stored and processed on third-party premises and in a shared multi-tenant environment; therefore, security and privacy vulnerabilities are increased. Providing an adequate solution is difficult as it needs to include both the service provider and the service consumer (Grolinger, 2013).

[148] Clojure is a dialect of the Lisp programming language. It is a general-purpose programming language with an emphasis on functional programming and runs on the Java Virtual Machine, Common Language Runtime and JavaScript engines.
[149] http://hpi.de/en/meinel/security-tech/cloud-security/cloudraid.html

5.3.8.2 Challenges and barriers

The following table shows barriers and challenges that arise within this topic:

Threat	Description/ Example
Data breaches	A data breach is defined as any incident involving the loss or exposure of digital personal records (Howard & Gulyas, 2014). As an example, in 2013 the retail store chain Target admitted that up to 110 million customer payment cards were compromised in a data breach that occurred in the Thanksgiving shopping period.
Data loss	Data stored in the cloud can be lost due to technical reasons such as accidental deletion by the cloud service provider, or a physical catastrophe such as a fire or earthquake, or if a company has encrypted data and loses the encryption key. Example: In 2009, for example, Health Net Insurance lost a portable hard drive containing financial and medical information of 1.5 million customers in the US.
Account hijacking	Account and service hijacking involves phishing, fraud and software vulnerabilities where attackers steal credentials and gain unauthorized access to servers (Khalil, Khreishah, & Azeem, 2014). In April 2010, for instance, Amazon experienced a Cross-Site Scripting (XSS) bug that allowed hackers to hijack credentials from the site.
Insecure APIs	From authentication and access control to encryption and activity monitoring, these interfaces must be designed to protect against both accidental and malicious attempts to circumvent policy.
Account misuse	In data driven and distributed environments, providers often require different approaches for handling security. Simple security errors such as too simple passwords or too similar passwords often lead to big negative impacts by accounts being taken over or misused.

Figure 6 shows the increased demand for secure systems and the increased significance within today's environments.

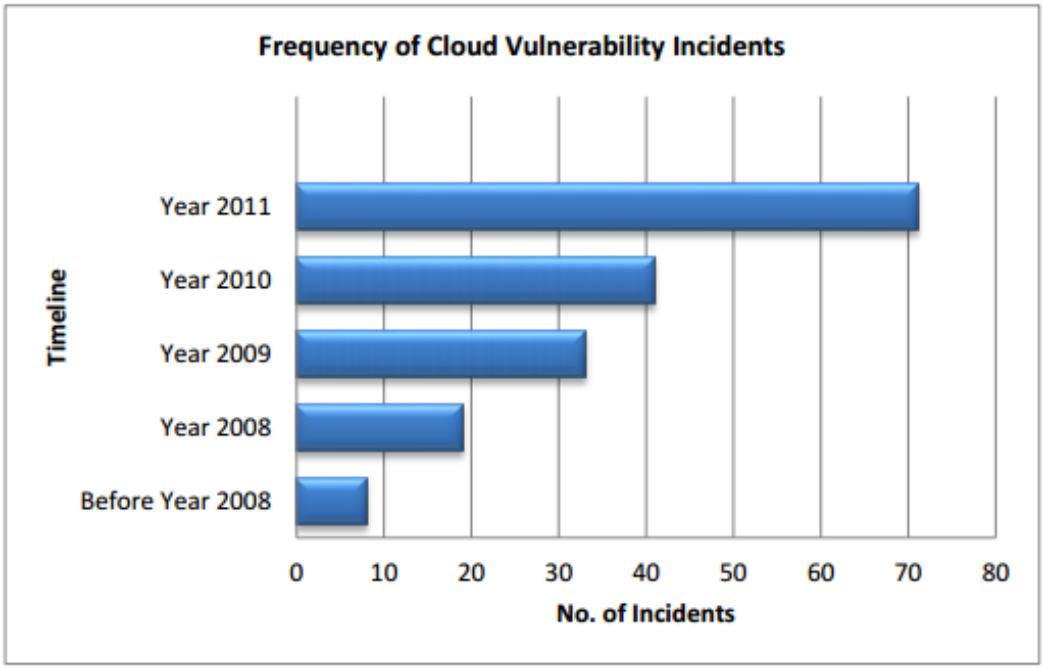

Figure 6: Frequency of cloud vulnerability incidents (Cloud Security Alliance, 2012)

5.3.8.3 Approaches and example cases

Data transferred between two of more partners on the web is often encrypted. Within web-based scenarios, HTTPS (Hypertext Transfer Protocol Secure) has been established as a standard for secure data exchange although it is still not used in all sensitive scenarios.

Web-based providers such as CloudFlare provide out-of-the-box protection against certain attacks including XSS and DOS (denial of service) attacks[150].

In order to prevent the misuse of accounts, several approaches have been developed which are relevant from a technological perspective. A very wide-spread example is OAuth (Open Authorization). OAuth is an authorization standard that enables users to grant third party applications with limited access to their resources stored on a server, without divulging their password or other secret credentials. Introduced in 2009, OAuth has gained widespread acceptance in a short period of time and can be called as de-facto standard for authorization. (Pai et al., 2011).

[150] https://www.cloudflare.com/

5.4 Summary and conclusion

From a technological perspective, the European data market is a highly diverse domain with a range of competing standards and de-facto standards. Certain trends such as JSON or NoSQL data storage mechanisms may, however, be identified.

Technology in the uprising data economy is currently not dominated by one major actor. It should, however, be mentioned that all major actors in this market are currently located in the US. The approach of realizing a European data economy may provide huge potentials within and beyond Europe as it allows organizations to better use and reuse information.

From a technological perspective, major challenges in this domain include the data management and data exchange as well as the interpretation and analysis of data which is often unsharp or uncertain.

6 Discussion

This section discusses the legal, socio-economic and technological analyses of framework conditions for both data reuse and the development of a self-sustaining data economy in Europe in an integrative manner.

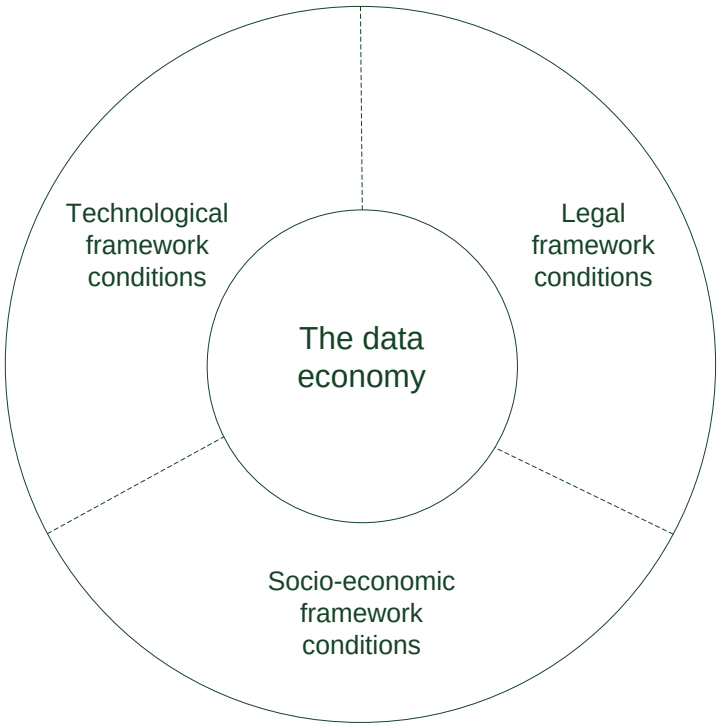

Figure 7: Framework conditions relevant for the data economy

Legal, socio-economic and technological framework conditions determine how the data economy evolves in Europe. Within the scope of EuDEco, it appeared reasonable to distinguish roughly between the previously named perspectives, the framework conditions, however, are partly overlapping and closely linked to each other. As shown in Figure 7, there are obvious touching points between the legal, the socio-economic and the technological perspective. The legal frameworks, for instance, is supposed to reflect the norms and value of the society but also it has to take the principles of economic activity into account. The institutional landscape which is discussed as part of the socioeconomic perspective defines policies and strategies and makes investments but also acts as legislator. The technological infrastructure and the economic actors' technology readiness are relevant from the socioeconomic-perspective but also from a technological one. Links between the technological and the legal perspective can also be identified easily. Technology developers and integrators have to take legal aspect into account, particularly as technologies tend to be created locally but used globally.

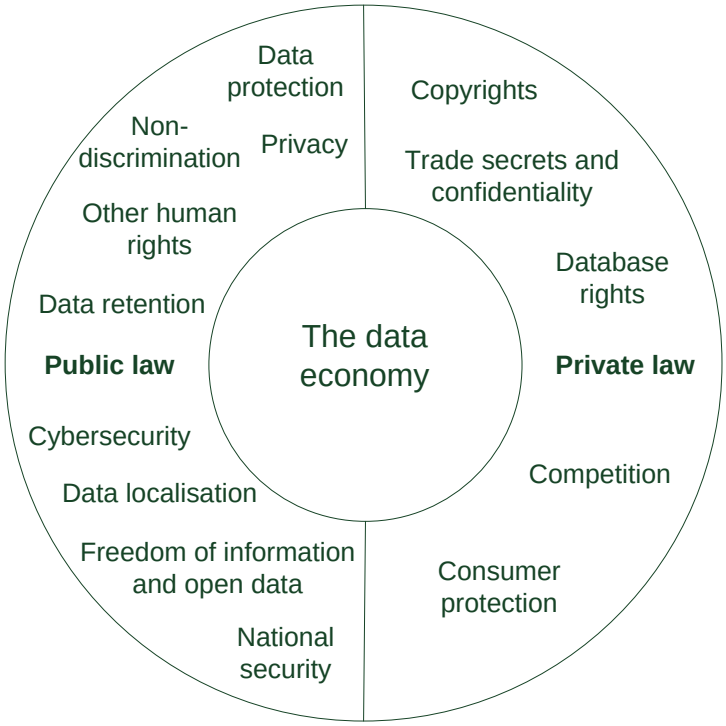

Figure 8: Legal framework conditions

From a legal perspective, as shown in Figure 8, framework conditions result from public law and private law. In principle, laws and regulations can be both barriers to and enablers of data reuse and the development of the data economy. With respect to public law only data retention law is considered a clear enabler as it ensures that data is available for additional analysis. Data protection law and cybersecurity can be considered to act as both barriers and enables, depending on the point of view. Data protection law, for instance, on the one hand aims at minimizing the amount of data and prevents

limitless retention, whereas, on the other hand, through he principle of data quality it leads to more precise datasets and in consequence to more reliable result. With respect to private law, all laws are considered to act as both barriers and enablers or data reuse and the data economy. This not surprising as the respective laws regulate relationships between actors of the data economy. The situation becomes quite obvious when looking at the example of IPR law which includes copyrights, database rights and trade secrets and confidentiality. It acts as a barrier when it limits data reusers to fully exploit datasets but as an enabler when it guarantees legal protection and thus encourages data creators to share their data.

Based on the current results, it will investigated how more data reuse may take place within the existing legal frameworks and how the legal framework may/should be modified in order to foster the exchange of data assets. Additionally, the role of legal certainty in that process and how the law can help setting up pilots for a self-sustaining data market will be investigated.

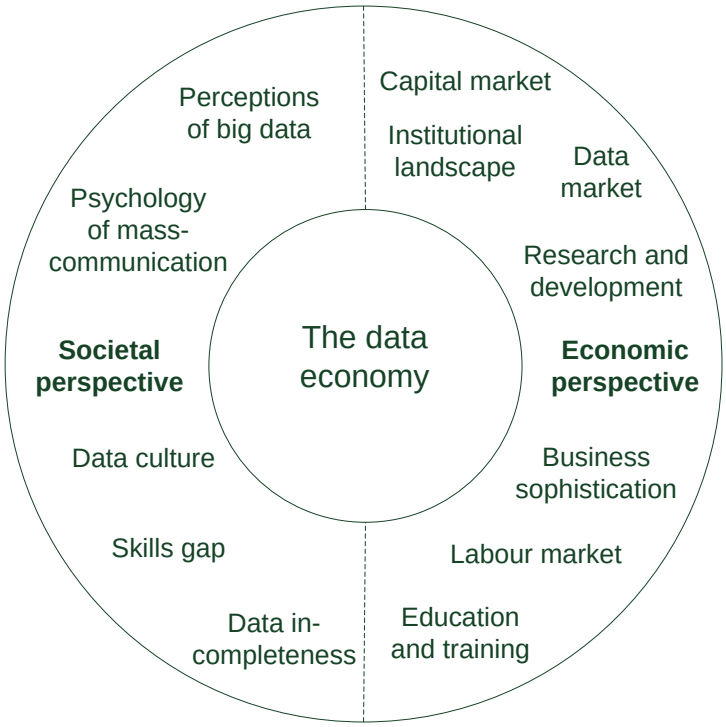

Figure 9: Socio-economic framework conditions

Aspects from which framework conditions originate that are relevant from a societal or an economic perspective are mentioned in Figure 9. There is a close relationship between social and economic framework conditions.

Based on the current results, it will be investigated how more data reuse may take place with the existing socio-economic framework conditions and how the framework conditions may/should be modified in order to foster the exchange of data assets. Moreover, it will be investigated in more detail how Europe differs from other regions in terms of socio-economic framework conditions and how the European data economy can be made improved. Improvement means taking the needs of both economy and society into account.

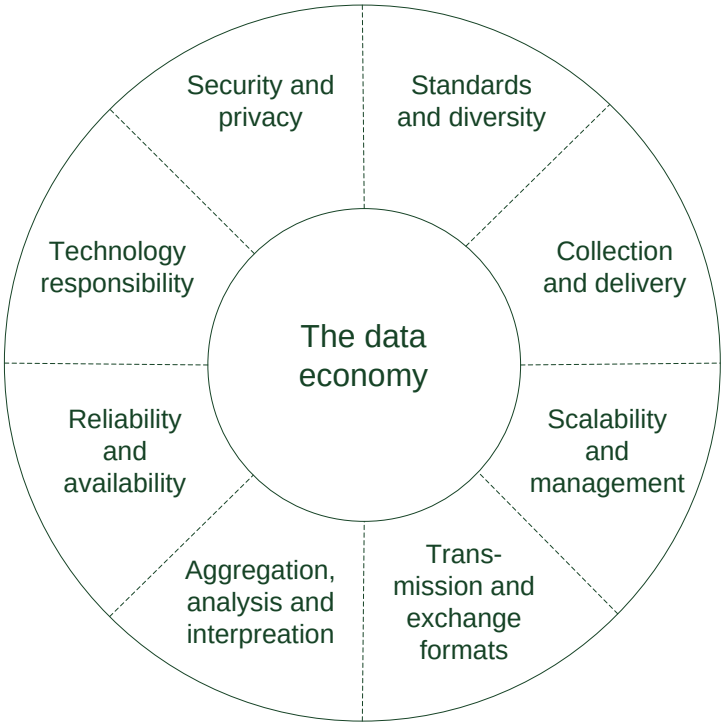

Figure 10: Technological framework conditions

Figure 10 provides an overview of aspects from which technological framework conditions originate. With respect to each aspect, challenges and barriers as well as approaches that may be used to overcome them could be identified. This reflects that, from a technological perspective, using and reusing large amounts of data is possible but that there is room for improvement in many areas. From a technological perspective, the European data market is a highly diverse domain with a range of competing standards and de-facto standards. All major actors in this market, however, are currently located in the US. The approach of realizing a European data economy may provide huge potentials within and beyond Europe as it allows organizations to better use and reuse information.

Based on the current results, it will be investigated how more data reuse may take place with the existing technological framework conditions and how the technologies may/should be modified in order to foster the exchange of data assets. Moreover, it will be investigated how Europe can move from a

user of technology to a provider of competitive technology as well as what advantages Europe would have if Europe succeeds in this.

7 Conclusions

A thorough understanding of the relevant framework conditions is essential to understand the data economy. The insights from the research that led to this document will directly flow into the initial, heuristic model of the data economy. The results presented in this document reflect the current status of research and the basis for the first model but are not to be considered final. They provide a first understanding and allow setting priorities for future research. While some of the aspects addressed in this document will be further investigated, others will not receive much more attention.

8 References

Acemoglu, D., Johnson, S., & Robinson, J. (2005). Institutions as the Fundamental Cause of Long-Run Growth. In P. Aghion & S. N. Durlauf (Eds.), *Handbooks in economics: Vol. 22. Handbook of economic growth* . Amsterdam: Elsevier.

Acquisti, A., & Gross, R. (2006). Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In *Proceedings of the Privacy Enhancing Technologies Workshop* .

Aeppel, T. (2014). *Bringing Jobs Back to U.S. Is Bruising Task: Stanley Furniture, Chesapeake Bay Candle Struggle With Their U.S.-Based Factories*. Retrieved from http://www.wsj.com/articles/bringing-jobs-back-to-u-s-is-bruising-task-1403746208?mod=ST1

Agrawal, D., El Abbadi, A., Antony, S., & Das, S. (2010). Data Management Challenges in Cloud Computing Infrastructures. In S. Kikuchi, S. Sachdeva, & S. Bhalla (Eds.), *Proceedings of the 6th International Workshop on Databases in Networked Information Systems (DNIS 2010)* (pp. 1–10).

Allan, R. (2.010). *Management and Analysis of Large Scientic Data Sets*. Retrieved from http://www.grids.ac.uk/NWGrid/LargeData.pdf

Armbrust, M., Stoica, I., Zaharia, M., Fox, A., Griffith, R., Joseph, A. D., . . . Rabkin, A. (2010). A view of cloud computing. *Communications of the ACM, 53*(4), 50. doi:10.1145/1721654.1721672

Article 29 Working Party. (2007). *Opinion 4/2007 on the concept of personal data* (No. WP136). Retrieved from http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf

Article 29 Working Party. (2008). *Opinion 1/2008 on data protection issues related to search engines* (No. WP148). Retrieved from http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2008/wp148_en.pdf

Article 29 Working Party. (2010). *Opinion 3/2010 on the principle of accountability* (No. WP173). Retrieved from http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp173_en.pdf

Article 29 Working Party. (2013). *Opinion 03/2013 on purpose limitation* (No. WP203). Retrieved from http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

Article 29 Working Party. (2014). *Opinion 05/2014 on Anonymisation Techniques* (No. WP216). Retrieved from http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Atkinson, R. D. & Mayo, M. (2010). *Refueling the U.S. Innovation Economy: Fresh Approaches to STEM Education*. Retrieved from http://www.itif.org/files/2010-refueling-innovation-economy.pdf

Bapat, A. (2013). The new right to data portability. *Privacy and Data Protection, 13*(3), 3–4.

Barlow, J. P. (1996). *A Declaration of the Independence of Cyberspace*. Retrieved from https://projects.eff.org/~barlow/Declaration-Final.html

Barth, M., & Fay, A. (2010). Efficient use of data exchange formats in engineering projects by means of language integrated queries — Engineers LINQ to XML. In *Proceedings of IECON 2010. 36th Annual Conference on IEEE Industrial Electronics Society* (pp. 1335–1340). Piscataway, NJ, USA: IEEE.

Bauer, M., Lee-Makiyama, H., Marel, E. van der, & Verschelde, B. (2014). *The Costs of Data Localisation: Friedly Fire on Economic Recovery* (ECIPE Occasional Paper No. 3/2014). Retrieved from http://www.ecipe.org/app/uploads/2014/12/OCC32014__1.pdf

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American, 284*(5), 28–37.

Bertolucci, J. (2013). *Big Data Analytics: Descriptive Vs. Predictive Vs. Prescriptive*. Retrieved from http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279

Black, H. C. (1910). *A Law Dictionary: Definitions of the terms and phrases of American and English Jurisprudence, ancient and modern* (2nd ed.): West Publishing Company.

Bollier, D. (2010). *The Promise and Peril of Big Data*. Retrieved from http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf

Boom, F. van den & Canova, L. (2013). *The most known challenges of PSI Access and Re-USE: Intellectual Property (and Data Protection)*. Retrieved from http://www.w3.org/2013/share-psi/workshop/lisbon/FreyjavandenBoom

Bosworth, D., Lyonette, C., Wilson, R., Bayliss, M., & Fathers, S. (2013). *The Supply of and Demand for High-Level STEM Skills: Evidence Report 77*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/302973/evidence-report-77-high-level-stem-skills_1_.pdf

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Bradley, J., Reberger, C., Dixit, A., & Gupta, V. (2013). *Internet of Everything: A $4.6 Trillion Public-Sector Opportunity: More Relevant, Valuable Connections Will Boost Productivity, Revenue, and Citizen Experience, While Cutting Costs*. Retrieved from http://internetofeverything.cisco.com/sites/default/files/docs/en/ioe_public_sector_vas_white%20paper_121913final.pdf

Bray, T. (2014). *The JavaScript Object Notation (JSON) Data Interchange Format: Request for Comments: 7159*. Retrieved from http://tools.ietf.org/html/rfc7159.html

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., & Cowan, J. (2006). *Extensible Markup Language (XML) 1.1 (Second Edition): W3C Recommendation*. Retrieved from http://www.w3.org/TR/2006/REC-xml11-20060816/

Brockhoff, J., Jehanno, B., Pozzato, V., Buhr, C.-C., Eberl, P., & Papandropoulos, P. (2008). Google/DoubleClick: The first test for the Commission's nonhorizontal merger guidelines. *Competition Policy Newsletter*, (2), 53–60.

Bygrave, L. A. (2000). European Data Protection: Determining Applicable Law pursuant to European Data Protection Legislation. *Computer Law & Security Review, 16*(4), 252–257. doi:10.1016/S0267-3649(00)89134-7

Bygrave, L. A. (2002). *Data protection law: Approaching its rationale, logic and limits*. *Information law series: Vol. 10*. The Hague: Kluwer Law International.

Capgemini Consulting. (2013). *The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data*. Retrieved from https://www.capgemini-consulting.com/resource-file-access/resource/pdf/opendata_pov_6feb.pdf

Chander, A., & Lê, U. P. (2014). Data nationalism. *Emory Law Journal, 64*(3). Retrieved from http://law.emory.edu/elj/_documents/volumes/64/3/articles/chander-le.pdf

Chen, S., Georg, F., & Loewi, A. (2011). *The Impact of Tablets on Information Availability*. Retrieved from http://cs.stanford.edu/people/eroberts/cs181/projects/2010-11/TabletDevices/index.html

Chiou, L. & Tucker, C. (2014). *Search Engines and Data Retention: Implications for Privacy and Antitrust* (MIT Sloan Research Paper No. 5094-14). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2441333##

Choi, T. Y., Dooley, K. J., & Rungtusanatham, M. (2001). Supply networks and complex adaptive systems: control versus emergence. *Journal of Operations Management, 19*(3), 351–366. doi:10.1016/S0272-6963(00)00068-1

Clark, T. (2015). *Can Big Data Survive Big Brother Hysteria?* Retrieved from http://www.forbes.com/sites/sap/2015/06/30/can-big-data-survive-big-brother-hysteria/

Cloud Security Alliance. (2012). *Cloud Computing Vulnerability Incidents: A Statistical Overview*. Retrieved from http://www.cert.uy/wps/wcm/connect/975494804fdf89eaabbdab1805790cc9/Cloud_Computing_Vulnerability_Incidents.pdf

Collin, M. (2014). *The difference between raising early-stage capital in the US vs. Europe: A founder's perspective*. Retrieved from http://tech.eu/features/3064/difference-raising-seed-capital-us-europe/

Competition & Markets Authority (CMA). (2015). *The commercial use of consumer data: Report on the CMA's call for information*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/435817/The_commercial_use_of_consumer_data.pdf

Consultative Committee for Space Data Systems. (2012). *Reference Model for an Open Archival Information System (OAIS): Recommendation for Space Data System Practices* (No. CCSDS 650.0-M-2). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Cook, T. (2010). *EU intellectual property law*. Oxford: Oxford Univ. Press.

Cornish, W., Llewelyn, D., & Aplin, T. (2013). *Intellectual property: Patents, copyright, trade marks and allied rights* (8. ed.). *Classics series*. London: Sweet & Maxwell.

Correa, T., & Jeong, S. H. (2011). Race and online content creation: Why minorities are actively participating in the Web. *Information, Communication & Society, 14*(5), 638–659. doi:10.1080/1369118X.2010.514355

Council of Canadian Academies. (2015). *Some Assembly Required: STEM Skills and Canada's Economic Productivity: The Expert Panel on STEM Skills for the Future*. Retrieved from http://www.scienceadvice.ca/uploads/ENG/AssessmentsPublicationsNewsReleases/STEM/STEMFullReportEn.pdf

Craig Richard. (2014). *Big Data and competition – merger control is not the remedy for data protection issues*. Retrieved from http://www.taylorwessing.com/globaldatahub/article_big_data_competition.html

Currie, D. (2013). *The new Competition and Markets Authority: how will it promote competition?* Beesley Lectures speech. Retrieved from https://www.gov.uk/government/speeches/the-new-competition-and-markets-authority-how-will-it-promote-competition

Custers, B. (2004). *The Power of Knowledge: Ethical, Legal, and Technological Aspects of Data Mining and Group Profiling in Epidemiology*: Wolf Legal Publishers.

Custers, B., Calders, T., Schermer, B., & Zarsky, T. (Eds.). (2013). *Studies in applied philosophy, epistemology and rational ethics: Vol. 3*. *Discrimination and privacy in the Information Society: Data mining and profiling in large databases*. Berlin, Heidelberg: Springer.

Custers, B., van der Hof, Simone, & Schermer, B. (2014). Privacy Expectations of Social Media Users: The Role of Informed Consent in Privacy Policies. *Policy & Internet, 6*(3), 268–295. doi:10.1002/1944-2866.POI366

Czapracka, K. A. (2007). Antitrust and Trade Secrets: The U.S. and the EU Approach. *Santa Clara High Technology Law Journal, 24*(2), 207–273. Retrieved from http://digitalcommons.law.scu.edu/chtlj/vol24/iss2/1

Davenport, T. H. & Dyché, J. (2013). *Big Data in Big Companies*. Retrieved from http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/bigdata-bigcompanies-106461.pdf

Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review, 90,* 70–76.

de Vries, M. & Hittmair, G. (2013). *Open data is fact now- when does the reuse start?* Retrieved from https://www.w3.org/2013/share-psi/wiki/images/a/a3/Lisbon_deVries_hittmair.pdf

Dean, J., & Ghemawat, S. (2010). MapReduce. *Communications of the ACM, 53*(1), 72. doi:10.1145/1629175.1629198

Dekkers, M., Polman, F., te Velde, R., & de Vries, M. (2006). *MEPSIR: Measuring European Public Secotr Information Ressources.*

Derclaye, E. (2008). *The legal protection of databases: A comparative analysis*. Cheltenham, UK, Northampton, MA: Edward Elgar.

DIGITALEUROPE. (2014). *Making Europe Fit for the Data Economy*. Retrieved from http://www.digitaleurope.org/DesktopModules/Bring2mind/DMX/Download.aspx?Command=Core_Download&entryID=864&PortalId=0&TabId=353

DLA Piper. (2015). *Data Protection Laws of the World*. Retrieved from http://dlapiperdataprotection.com/system/modules/za.co.heliosdesign.dla.lotw/functions/export.pdf?country=all

Donvito, G., Marzulli, G., & Diacono, D. (2014). Testing of several distributed file-systems (HDFS, Ceph and GlusterFS) for supporting the HEP experiments analysis. *Journal of Physics: Conference Series, 513*(4). doi:10.1088/1742-6596/513/4/042014

Drukker Solicitors. (2014). *The Widening Girth of Legal Protection for Confidential Information*. Retrieved from http://www.drukker.co.uk/publications/l/the-widening-girth-of-legal-protection-for-confidential-information/#.VcCIPMsw8Ud

Dulong de Rosnay, M., & Janssen, K. (2014). Legal and Institutional Challenges for Opening Data across Public Sectors: Towards Common Policy Solutions. *Journal of theoretical and applied electronic commerce research, 9*(3), 1–14. doi:10.4067/S0718-18762014000300002

Dutton, W. H. & Blank, G. (2013). *Cultures of the Internet: The Internet in Britain: Oxford Internet Survey 2013 Report*. Retrieved from http://oxis.oii.ox.ac.uk/wp-content/uploads/2014/11/OxIS-2013.pdf

e-skills UK. (2013). *Big Data Analytics: An assessment of demand for labour and skills, 2012-2017*. Retrieved from http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=6243

European Commission. (2005). *First evaluation of Directive 96/9/EC on the legal protection of databases*. Retrieved from http://ec.europa.eu/internal_market/copyright/docs/databases/evaluation_report_en.pdf

European Commission. (2010). *A Digital Agenda for Europe: COM(2010) 245*. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52010DC0245&from=EN

European Commission. (2011a). *Open data: An engine for innovation, growth and transparent governance: COM(2011)882 final*. Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF

European Commission. (2011b). *Special Eurobarometer 359: Attitudes on Data Protection and Electronic Identity in the European Union*. Retrieved from http://ec.europa.eu/public_opinion/archives/ebs/ebs_359_en.pdf

European Commission. (2012). *Data protection reform: Frequently asked questions*. Retrieved from http://europa.eu/rapid/press-release_MEMO-12-41_de.htm?locale=en

European Commission. (2013a). *Proposal for a Directive of the European parliament and of the Council concerning measures to ensure a high common level of network and information security across the Union* (No. COM(2013) 48 final).

European Commission. (2013b). *Study on Trade Secrets and Confidential Business Information in the Internal Market*. Retrieved from http://ec.europa.eu/internal_market/iprenforcement/docs/trade-secrets/130711_final-study_en.pdf

European Commission. (2014a). *Factsheet on the 'Right to be forgotten' ruling: C-131/12*. Retrieved from http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_en.pdf

European Commission. (2014b). *Guidelines on recommended standard licences, datasets and charging for the reuse of documents*. Retrieved from https://ec.europa.eu/digital-agenda/en/news/commission-notice-guidelines-recommended-standard-licences-datasets-and-charging-re-use

European Commission. (2014c). *Towards a thriving data-driven economy: COM(2014)442 final*. Retrieved from http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=6210

European Commission. (2015). *A Digital Single Market Strategy for Europe* (No. COM(2015) 192 final). Retrieved from http://ec.europa.eu/priorities/digital-single-market/docs/dsm-communication_en.pdf

European Council. (2013). *Conclusions: EUCO 169/13*. Retrieved from http://www.consilium.europa.eu/uedocs/cms_Data/docs/pressdata/en/ec/139197.pdf

European Data Protection Supervisor. (2014a). *Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy*. Brussels. Retrieved from https://secure.edps.europa.eu/EDPSWEB/webdav/shared/Documents/Consultation/Opinions/2014/14-03-26_competitition_law_big_data_EN.pdf

European Data Protection Supervisor. (2014b). *Report of workshop on Privacy, Consumers, Competition and Big Data 2 June*. Retrieved from https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Big%20data/14-07-11_EDPS_Report_Workshop_Big_data_EN.pdf

European Union. *Recommendation of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning: 2006/962/EC* (Official Journal of the European Union). Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006H0962&from=EN

European Union Agency for Fundamental Rights & Council of Europe. (2014). *Handbook on European data protection law*. Luxembourg. Retrieved from http://fra.europa.eu/sites/default/files/fra-2014-handbook-data-protection-law-2nd-ed_en.pdf

Executive Office of the President. (2014). *Big Data: Seizing Opportunities, Preserving Values*. Retrieved from https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

EY. (2014). *(Big) data - Où en sont les entreprises françaises?: Quelle maturité dans l'exploitation des données clients?* Retrieved from http://www.ey.com/FR/fr/Services/Advisory/EY-etude-Big-data-2014

Fabernovel. (2014). *GAFAnomics: New economy, new rules*. Retrieved from http://www.fabernovel.com/work/study-gafanomics-new-economy-new-rules/

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data Analysis. *National science review, 1*(2), 293–314. doi:10.1093/nsr/nwt032

Fecher, B., Friesike, S., Hebing, M., Linek, S., & Sauermann, A. (2015). *A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing* (Discussion Papers No. 1454). Retrieved from http://www.diw.de/documents/publikationen/73/diw_01.c.497416.de/dp1454.pdf

Federal Trade Commission. (2014). *Data Brokers: A Call for Transparency and Accountability*. Retrieved from https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

Feuerlicht, G. (2010). Database Trends and Directions: Current Challenges and Opportunities. In J. Pokorný, V. Snášel, & K. Richta (Eds.), *Proceedings of the Dateso 2010 Workshop. Databases, Texts, Specifications, and Objects* .

Filippi, P. de, & Gracz, K. (2013). Resolving the Crisis of Copyright Law in the Digital Environment: Reforming the 'Copy-Right' into a 'Reuse-Right'. In *Proceedings of the 7th International Conference on the Interaction of Knowledge Rights, Data Protection and Communication* .

Franks, B. (2012). *Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics*. *Wiley & SAS business series*. Hoboken, New Jersey: John Wiley & Sons Inc.

Fujitsu. (2013). *Solving the Big Dilemma of Big Data*. Retrieved from http://www.fujitsu.com/us/Images/BigDatawp.pdf

Gamma, E., Helm, R., Johnson, R. E., & Vlissides, J. (1994). *Design patterns: Elements of reusable object-oriented software*: Prentice Hall.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144. doi:10.1016/j.ijinfomgt.2014.10.007

Gartner. (2013). *Gartner Identifies the Top 10 Strategic Technology Trends for 2013*. Retrieved from http://www.gartner.com/newsroom/id/2209615

Gartner. (2014). *Gartner Identifies the Top 10 Strategic Technology Trends for 2014*. Retrieved from http://www.gartner.com/newsroom/id/2603623

Geradin, D. (2014). *Data portability and EU competition law*. Retrieved from http://www.cullen-international.com/asset/?location=/content/assets/training--conferences/conferences/2014/04/damien-geradin---data-portability-and-eu-competition-law.pdf/damien-geradin---data-portability-and-eu-competition-law.pdf

Geradin, D. & Kuschewsky, M. (2013). *Competition Law and Personal Data: Preliminary Thoughts on a Complex Issue*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2216088

Giannotti, F., Nanni, M., Rauber, A., & Thanos, C. (2014). *Big Data Analytics: towards a European research agenda*. Retrieved from https://europe.rd-alliance.org/sites/default/files/BigDataAnalytics_white_paper_v9_0.pdf

Glover, M., Holsen, S., MacDonald, C., Rahman, M., & Simpson, D. (2006). *Freedom of Information: History, Experience and Records and Information Management Implications in the USA, Canada and the United Kingdom*. Retrieved from http://www.armaedfoundation.org/pdfs/Freedom_of_Information_in_US_UK_and_Canada.pdf

Graham, R. & Lewington, A. (2013). *The Big Data Explosion: A New Frontier in Digital Law*. Retrieved from http://www.scl.org/site.aspx?i=ed31114

Graves, A., & Bustos-Jiménez, J. (2014). Towards Visual Overviews for Open Government Data. In *Proceedings of Data Wiz 2014* .

Griffith, R., Virmani, R., & Kaiser, G. E. (2007). *The Role of Reliability, Availability and Serviceability (RAS) Models in the Design and Evaluation of Self-Healing Systems* (Columbia University Academic Commons). Retrieved from http://hdl.handle.net/10022/AC:P:29507

Grolinger, K. (2013). *Disaster Data Management in Cloud Environments* (Electronic Thesis and Disseration Repository No. 1774).

Guo, C., Lu, G., Li, D., Wu, H., Zhang, X., Shi, Y., . . . Lu, S. (2009). BCube: a high performance, server-centric network architecture for modular data centers. *ACM SIGCOMM Computer Communication Review, 39*(4), 63. doi:10.1145/1594977.1592577

Gurin, J. (2014). *Big data and open data: what's what and why does it matter?* Retrieved from http://www.theguardian.com/public-leaders-network/2014/apr/15/big-data-open-data-transform-government

Hage, J., & Akkermans, B. (Eds.). (2014). *Introduction to Law*. Cham, s.l.: Springer International Publishing.

Hahnel, M. (2015). *What's the point of open academic data? Who will it affect? - See more at: http://figshare.com/blog/What_s_the_point_of_open_academic_data_Who_will_it_affect_/144#sthash.3S0VZdCn.dpuf*. Retrieved from http://figshare.com/blog/What_s_the_point_of_open_academic_data_Who_will_it_affect_/144

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47,* 98–115. doi:10.1016/j.is.2014.07.006

Hasselbring, W., Reussner, R., Jaekel, H., Schlegelmilch, J., Teschke, T., & Krieghoff, S. (2004). The Dublo Architecture Pattern for Smooth Migration of Business Information Systems: An Experience Report. In *Proceedings of the 26th International Conference on Software Engineering (ICSE 2004)* (pp. 117–126). IEEE Computer Society.

Hill, J. F. (2014). *The Growth of Data Localization Post-Snowden: Analysis and Recommendations for U.S. Policymakers and Business Leaders*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2430275

Hoboken, J. van. (2013). *The Proposed Right to be Forgotten Seen from the Perspective of Our Right to Remember: Freedom of Expression Safeguards in a Converging Information Environment*. Retrieved from http://www.law.nyu.edu/sites/default/files/upload_documents/VanHoboken_RightTo%20Be%20Forgotten_Manuscript_2013.pdf

Holt, R. C., Schürr, A., Sim, S. E., & Winter, A. (2006). GXL: A graph-based standard exchange format for reengineering. *Science of Computer Programming, 60*(2), 149–170. doi:10.1016/j.scico.2005.10.003

Howard, P. N. & Gulyas, O. (2014). *Data Breaches in Europe: Reported Breaches of Compromised Personal Records in Europe, 2005-2014*. Retrieved from http://cmds.ceu.edu/article/2014-10-07/data-breaches-europe-reported-breaches-compromised-personal-records-europe-2005

Hsu, L., Martin, R. L., McElroy, B., Litwin-Miller, K., & Kim, W. (2015). Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities. *Geomorphology, 244,* 180–189. doi:10.1016/j.geomorph.2015.03.039

Hunton & Williams. (2015). *Council of the European Union Agrees on General Approach to the Proposed General Data Protection Regulation*. Retrieved from https://www.huntonprivacyblog.com/2015/06/16/council-european-union-agrees-general-approach-proposed-general-data-protection-regulation/

Information Commissioner's Office (ICO). (n.d.). *Key definitions of the Data Protection Act*. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/

Information Commissioner's Office (ICO). (2014). *Big data and data protection*. Retrieved from https://ico.org.uk/media/for-organisations/documents/1541/big-data-and-data-protection.pdf

Janssen, K., & Dumortier, J. (2007). Legal framework for a European Union Spatial Data Infrastructure: Uncrossing the Wires. In H. Onsrud (Ed.), *Research and Theory in Advancing Spatial Data Infrastructure Concepts* . Redlands, CA: ESRI Press. Retrieved from http://www.gsdidocs.org/gsdiconf/GSDI-9/papers/TS8.1paper.pdf

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. In *Proceedings of the 46th Hawaii International Conference on System Sciences (HICSS)* (pp. 995–1004). IEEE.

Kalas, M., Puntervoll, P., Joseph, A., Bartaseviciute, E., Topfer, A., Venkataraman, P., . . . Jonassen, I. (2010). BioXSD: the common data-exchange format for everyday bioinformatics web services. *Bioinformatics, 26*(18), i540–i546. doi:10.1093/bioinformatics/btq391

Kalyvas, J. R., & Overly, M. R. (2014). *Big data: A business and legal guide*. *An Auerbach Book*. Boca Raton, Fla.: CRC Press.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing, 74*(7), 2561–2573. doi:10.1016/j.jpdc.2014.01.003

Kamiran, F., & Calders, T. (2009). Classification without discrimination. In *Proceedings of the IEEE International Conference on Computer, Control and Communication (IEEE-IC4)* . IEEE.

Katti, S., Katabi, D., Hu, W., Rahul, H., & Médard, M. (2006). The Importance of Being Opportunistic: Practical Network Coding for Wireless Environments. *ACM SIGCOMM Computer Communication Review, 36*(4).

Kemp, R. (2013). *Big Data: Legal Rights and Obligations*. Retrieved from http://www.kemplittle.com/cms/document/Big_Data_Legal_Rights_and_Obligations.pdf

Kemp, R. (2014). *Legal Aspects of Managing Big Data*. Retrieved from http://www.kempitlaw.com/wp-content/uploads/2014/10/Legal-Aspects-of-Big-Data-White-Paper-v2-1-October-2014.pdf

Khalil, I., Khreishah, A., & Azeem, M. (2014). Cloud Computing Security: A Survey. *Computers, 3*(1), 1–35. doi:10.3390/computers3010001

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, Waleed Kamaleldin, Alam, M., . . . Gani, A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal, 2014*(6077), 1–18. doi:10.1155/2014/712826

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography, 3*(3), 262–267. doi:10.1177/2043820613513388

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1). doi:10.1177/2053951714528481

Klyne, G., Carroll, J. J., & McBride, B. (2014). *RDF 1.1 Concepts and Abstract Syntax: W3C Recommendation*. Retrieved from http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

Kolaitis, P. G. (2005). Schema mappings, data exchange, and metadata management. In *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 61–75).

Korn, N. & Oppenheim, C. (2011). *Licensing Open Data: A Practical Guide*. Retrieved from http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802–5805. doi:10.1073/pnas.1218772110

Kroes, N. (2013). *Big data for Europe*. Speech at ICT 2013. Retrieved from http://europa.eu/rapid/press-release_SPEECH-13-893_en.htm

Kshetri, N. (2014). The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns. *Big Data & Society, 1*(2). doi:10.1177/2053951714564227

Kumar, R., Parashar, B. B., Gupta, S., Sharma, Y., & Gupta, N. (2014). Apache Hadoop, NoSQL and NewSQL Solutions of Big Data. *International Journal of Advance Foundation and Research in Science & Engineering, 1*(6), 28–36.

Kuner, C. (2007). *European data protection law: Corporate compliance and regulation* (2. ed.). New York: Oxford Univ. Press.

Kuner, C. (2012). The European Commission's Proposed Data Protection Regulation: A Copernican Revolution in European Data Protection Law. *Bloomberg BNA Privacy and Security Law Report,* 1–15.

Kuner, C., Cate, F. H., Millard, C., & Svantesson, D. J. B. (2012). The challenge of 'big data' for data protection. *International Data Privacy Law, 2*(2), 47–49. doi:10.1093/idpl/ips003

Kur, A., & Dreier, T. (2013). *European intellectual property law: Text, cases and materials*. Cheltenham: Elgar.

LAPSI. (n.d.a). *Policy Recommendation N. 1: The Competition Law Issues of the Re-Use of Public Sector Information (PSI).*

LAPSI. (n.d.b). *Policy Recommendation N. 2: The Interface Between the Protection of Commercial Secrecy and the Re-Use of Public Sector Information.*

LAPSI. (2012). *Policy Recommendation N. 4: Privacy and Personal Data Protection*. Retrieved from http://www.ivir.nl/publicaties/download/1098

Loshin, D. (2011). *The practitioner's guide to data quality improvement*. *The MK / OMG Press*. Amsterdam: Morgan Kaufmann/Elsevier.

Lynch, L. (2013). *EU Data Protection's Paradigm Shift: From Directive to Regulation*. Amsterdam. Retrieved from http://www.ey.com/Publication/vwLUAssets/EU_Data_Protections_Paradigm_Shift_From_Directive_to_Regulation/$FILE/EU%20Data%20Protections%20Paradigm%20Shift%20From%20Directive%20to%20Regulation.pdf

Mandel, M. (2014). *Where are the Big Data Jobs?* Retrieved from http://www.progressivepolicy.org/wp-content/uploads/2014/05/2014.05-Mandel_Where-Are-the-Big-Data-Jobs.pdf

Manovich, L. (2011). *Trending: The Promises and the Challenges of Big Social Data*. Retrieved from http://manovich.net/content/04-projects/067-trending-the-promises-and-the-challenges-of-big-social-data/64-article-2011.pdf

Mathae, K. B., & Uhlir, P. F. (Eds.) 2012. *The case for international sharing of scientific data: A focus on developing countries.: Proceedings of a Symposium*. Washington DC: National Academy of Sciences.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. *An Eamon Dolan book*. Boston: Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. *Harvard Business Review, 90*(10), 61–67.

McCallum, Q. E., & Gleason, K. (2013). *Business models for the data economy*: O'Reilly Media.

McKean, R. (2014). *EU Data Protection reform: where are we - and what can you do to prepare?* Retrieved from http://www.olswang.com/articles/2014/12/eudpreform/

McKinsey Global Institute. (2011). *Internet matters: The Net's sweeping impact on growth, jobs, and prosperity*. Retrieved from http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights and pubs/MGI/Research/Technology and Innovation/Internet matters - Nets sweeping impact/MGI_internet_matters_full_report.ashx

Miani, C., Robin, E., Horvath, V., Manville, C., Cave, J., & Chataway, J. (2014). *Health and Healthcare: Assessing the Real-World Data Policy Landscape in Europe*. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_reports/RR500/RR544/RAND_RR544.pdf

Miličić, V. (2011). *Problems of the RDF syntax*. Retrieved from http://milicicvuk.com/blog/2011/07/21/problems-of-the-rdf-syntax/

Moerel, E. M. L. (2011). *Binding corporate rules: Fixing the regulatory patchwork of data protection*. Tilburg: Tilburg Institute for Law, Technology, and Society.

Moyle, K., Blackall, L., Hudson, F., & Mason, J. (2012). *Open Data and Open Data and Open Standards.*

Muckersie, E. (2013). *Big Data: Big Opportunity or Big Brother?* Retrieved from http://www.freshminds.net/2013/03/big-data-big-opportunity-big-brother/

Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review, 79*(1), 119–157.

NTT DATA. (2015). *The EU's Big Data Strategy Starts Rolling*. Retrieved from http://www.nttdata.com/global/en/insights/it-briefings/2015032601.html

OECD. (2006). *Digital Broadband Content: Public Sector Information and Content*. Retrieved from http://www.oecd.org/sti/36481524.pdf

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review, 57,* 1701–1777.

Orwell, G. (1949). *1984*. London: Secker & Warburg.

Pai, S., Sharma, Y., Kumar, S., Pai, R. M., & Singh, S. (2011). Formal Verification of OAuth 2.0 Using Alloy Framework. In G. Tomar (Ed.), *Proeedings of the 2011 International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 655–659). Piscataway, NJ: IEEE.

Parameswaran, R. (2006). *A Robust Data Obfuscation Technique for Privacy Preserving Collaborative Filtering*. Retrieved from http://hdl.handle.net/1853/11459

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International conference on Knowledge discovery and data mining* (pp. 560–568).

Peers, S. (2015). *When super-regulators fight: the 'one-stop shop' in the proposed Data Protection Regulation*. Retrieved from http://eulawanalysis.blogspot.de/2015/03/when-super-regulators-fight-one-stop.html

Pfanner, E. (2013). *French Tax Proposal Zeroes In on Web Giants' Data Harvest*. Retrieved from http://www.nytimes.com/2013/02/25/technology/french-tax-proposal-zeroes-in-on-web-giants-data-harvest.html?_r=1

Pira International Ltd., University of East Anglia, & KnowledgeView Ltd. (2000). *Commercial exploitation of Europe's public sector information: Executive summary*. Retrieved from ftp://ftp.cordis.europa.eu/pub/econtent/docs/2000_1558_en.pdf

Pollock, R. (2009). *Why does this matter?* Retrieved from http://blog.okfn.org/2009/02/02/open-data-openness-and-licensing/

Price, B. A., Adam, K., & Nuseibeh, B. (2005). Keeping ubiquitous computing to yourself: A practical model for user control of privacy. *International Journal of Human-Computer Studies, 63*(1-2), 228–253. doi:10.1016/j.ijhcs.2005.04.008

Proust, O. (2015). *The EU General Data Protection Regulation is on its way…but when?* Retrieved from http://privacylawblog.fieldfisher.com/2015/the-eu-dp-regulation-is-on-its-way-but-when

Rampini, M. (2015). *How to develop a data culture?* Retrieved from http://www.forum-avignon.org/en/contribution-how-develop-data-culture-marion-rampini

Rauhofer, J., & Mac Sithigh, D. (2014). The Data Retention Directive Never Existed. *SCRIPTed, 11*(1).

Ravikumar, G. K., Manjunath, T. N., Ravindra, S., & Umesh, I. M. (2011). A Survey on Recent Trends, Process and Development in Data Masking for Testing. *International Journal of Computer Science Issues, 8*(2), 535–544.

RECODE. (2014). *Legal and ethical issues in open access and data dissemination and preservation: Deliverable D3.1*. Retrieved from http://recodeproject.eu/wp-content/uploads/2014/05/D3.1-legal-and-ethical-issues-FINAL.pdf

Regan, P. (2003). Privacy and Commercial Use of Personal Data: Policy Developments in the United States. *Journal of Contingencies and Crisis Management, 11*(1), 12–18. doi:10.1111/1468-5973.1101003

Reggi, L. (2011). *Open Data to the next level: WHY and HOW to involve the private sector*. Retrieved from http://luigireggi.eu/2011/09/19/open-data-to-the-next-level-why-and-how-to-involve-the-private-sector/

Ricolfi, M. (n.d.). *Concetural Framework No 2: Licensing: A Conceptual Framework for EU Guidance to the Member States.*

Robinson, D., Yu, H., Zeller, W. P., & Felten, E. W. (2009). Government Data and the Invisible Hand. *Yale Journal of Law and Technology, 11*(1). Retrieved from http://digitalcommons.law.yale.edu/yjolt/vol11/iss1/4

Rosen, J. (2012). The right to be forgotten. *Stanford Law Review Online, 64,* 88.

Rosenfeld, M. (2013). Rethinking the boundaries between public law and private law for the twenty first century: An introduction. *International Journal of Constitutional Law, 11*(1), 125–128. doi:10.1093/icon/mos053

Rubinstein, I. S. (2013). Big Data: The End of Privacy or a New Beginning? *International Data Privacy Law, 3*(2), 74–87. Retrieved from http://idpl.oxfordjournals.org/content/3/2/74.full.pdf

Ryan, T., & Xenos, S. (2011). Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior, 27*(5), 1658–1664. doi:10.1016/j.chb.2011.02.004

Salamanca, O. & Eechoud, M. van. (2014). *Open Legal Data for Europe: LAPSI/Openlaws Workshop*. Workshop Report. Retrieved from http://www.openlaws.eu/wp-content/uploads/2014-09-02_LAPSI-Openlaws_workshop-Amsterdam.pdf

Savvas, A. (2013). *IDC: Hadoop commonly used with other big data analytics systems: Enterprise Hadoop deployment survey reveals complex picture*. Retrieved from http://www.computerworlduk.com/news/data/idc-hadoop-commonly-used-with-other-big-data-analytics-systems-3476789/

Schermer, B. (2015). *Towards a general data law?* Retrieved from http://leidenlawblog.nl/articles/towards-a-general-data-law

Schermer, B. W., Custers, B., & van der Hof, Simone. (2014). The crisis of consent: how stronger legal protection may lead to weaker consent in data protection. *Ethics and Information Technology, 16*(2), 171–182. doi:10.1007/s10676-014-9343-8

Schomm, F., Stahl, F., & Vossen, G. (2013). Marketplaces for data: an initial survey. *ACM SIGMOD Record, 42*(1), 15. doi:10.1145/2481528.2481532

Scott, M. (2015). *For Tech Start-Ups in Europe, an Oceanic Divide in Funding*. Retrieved from http://www.nytimes.com/2015/02/14/technology/for-tech-start-ups-in-europe-an-oceanic-divide-in-funding.html?_r=0

Sekhar, J., Jeba, G., & Durga, S. (2012). A survey on energy efficient server consolidation through vm live migration. *International Journal of Advances in Engineering & Technology, 5*(1), 515–525.

Severance, C. (2012). Discove ring JavaScript Object Notation. *Computer, 45*(4), 6–8. doi:10.1109/MC.2012.132

Seybert, H. & Reinecke, P. (2013). *Internet use statistics - individuals: Three quarters of Europeans used the internet in 2013*. Retrieved from http://ec.europa.eu/eurostat/statistics-explained/index.php/Internet_use_statistics_-_individuals

Shadbolt, N. (2010). *Towards a pan EU data portal – data.gov.eu*. Retrieved from http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=1191

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. In *Proceedings of IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)* (pp. 1–10). IEEE.

Siewiorek, D. P., & Swarz, R. S. (2014). *Reliable Computer Systems: Design and Evaluation* (2nd). Burlington, MA, USA: Digital Press.

Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review, 154*(3), 477–560.

Solove, D. J. (2007). *The future of reputation: Gossip, rumor, and privacy on the internet*. New Haven Conn. u.a.: Yale Univ. Press.

Stern, M. J., Adams, Alison E., & Boase, J. (2011). Rural Community Participation, Social Networks, and Broadband Use: Examples from Localized and National Survey Data: . *Agricultural and Resource Economics Review, 40*(2), 158–171.

Stucke, M. E., & Grunes, A. P. (2015). Debunking the Myths Over Big Data and Antitrust. *CPI Antitrust Chronicle, 5*(2).

Swire, P., & Lagos, Y. (2013). Why the Right to Data Portability Likely Reduces Consumer Welfare: Antitrust and Privacy Critique. *Maryland Law Review, 72*(2).

Tassey, G. (1999). *Standardization in Technology-Based Markets*. Retrieved from http://www.nist.gov/director/planning/upload/researchpolicypaper.pdf

Tene, O., & Polonetsky, J. (2013). Big Data for All: Privacy and User Control in the Age of Analytics. *Northwestern Journal of Technology and Intellectual Property, 11*(5). Retrieved from http://scholarlycommons.law.northwestern.edu/njtip/vol11/iss5/1

Thalmann, S., Bachlechner, D., Demetz, L., & Maier, R. (2012). Challenges in Cross-Organizational Security Managemen t. In *Proceedings of the 45th Hawaii International Conference on System Sciences* (pp. 5480–5489).

The Australian Industry Group. (2015). *Progressing STEM Skills in Australia*. Retrieved from http://www.aigroup.com.au/portal/binary/com.epicentric.contentmanagement.servlet.ContentDeliveryServlet/LIVE_CONTENT/Publications/Reports/2015/14571_STEM%2520Skills%2520Report%2520Final%2520-.pdf

Tongia, R., Subrahmanian, E., & Arunachalam, V. S. (2005). *Information and Communcations Technology for Sustainable Development: Defining a Global Research Agenda*. Retrieved from https://www.andrew.cmu.edu/user/tongia/ICT4SD_final_draft-public.pdf

Tost, D. (2015). *German government repackages data retention regulations*. Retrieved from http://www.euractiv.com/sections/infosociety/german-government-repackages-data-retention-regulations-313821

Trivedi, N. (2014). *Codename Elastic Storage: A Software Defined Storage-as-a-Service Offering on SoftLayer for High Performance Data Management in Cloud*. Retrieved from https://www.ibm.com/developerworks/community/blogs/ibmsyssw/entry/codename_elastic_storage_a_software_defined_storage_as_a_service_offering_on_softlayer_for_high_performance_data_management_in_cloud?lang=en

Tucker, D. S., & Wellford, H. B. (2014). Big Mistakes Regarding Big Data. *The Antitrust Source, 14*(2). Retrieved from http://www.americanbar.org/content/dam/aba/publishing/antitrust_source/dec14_tucker_12_16f.authcheckdam.pdf

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics, 33*(1), 1–67. doi:10.1214/aoms/1177704711

UK Data Archive. (n.d.). *Research Data Lifecycle*. Retrieved from http://www.data-archive.ac.uk/create-manage/life-cycle

Ursic, H. (2015). The upcoming EU regulation on medical devices restates data protection principles. *Eye for Privacy,* 12.

van Alsenoy, B. & Koekkoek, M. (2015). *Internet and Jurisdiction after Google Spain: The extra-Territorial Reach of the EU's 'Right to be forgotten'* (No. 152). Retrieved from http://ghum.kuleuven.be/ggs/publications/working_papers/new_series/wp151-160/wp152-alsenoy-koekkoek.pdf

van Alsenoy, B., Verdoodt, V., Heyman, R., Wauters, E., Ausloos, J., & Acar, G. (2015). *From social media service to advertising network: a critical analysis of Facebook's Revised Policies and Terms*. Retrieved

from https://www.law.kuleuven.be/icri/en/news/item/facebooks-revised-policies-and-terms-v1-2.pdf

Vedaschi, A., & Lubello, V. (2015). Data Retention and its Implications for the Fundamental Right to Privacy. *Tilburg Law Review, 20,* 14–34.

Vedder, A. (1999). *Ethics and Information Technology, 1*(4), 275–281. doi:10.1023/A:1010016102284

Verheyen, S. (2015). *EU must update copyright to maintain leading creative sector*. Retrieved from https://www.theparliamentmagazine.eu/articles/opinion/eu-must-update-copyright-maintain-leading-creative-sector

Vickery, G. (2011). *Review of recent studies of PSI reuse and related market developments*. Retrieved from http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=1093

viEUws. (2015). *'Passenger name record needs to comply with ECJ ruling on data retention' S&D argues*. Retrieved from http://www.vieuws.eu/ict/passenger-name-record-needs-to-comply-with-ecj-ruling-on-data-retention-sd-argues/

Waller, M. A., & Fawcett, S. E. (2013). Click Here for a Data Scientist: Big Data, Predictive Analytics, and Theory Development in the Era of a Maker Movement Supply Chain. *Journal of Business Logistics, 34*(4), 249–252. doi:10.1111/jbl.12024

Wang, G., & Ng, T. S. E. (2010). The Impact of Virtualization on Network Performance of Amazon EC2 Data Center. In *Proceedings of INFOCOM 2010* (pp. 1–9). Piscataway, NJ, USA: IEEE.

Warren, S. D., & Brandeis, L. D. (1890). The Right to Privacy. *Harvard Law Review, 4*(5). Retrieved from http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html

Whish, R., & Bailey, D. (2012). *Competition law* (Seventh edition). New York: Oxford University Press.

White, T. E. (2012). *Hadoop: The definitive guide* (3rd ed.). Sebastopol, CA: O'Reilly.

Winborg, J., & Landström, H. (2001). Financial bootstrapping in small businesses. *Journal of Business Venturing, 16*(3), 235–254. doi:10.1016/S0883-9026(99)00055-5

World Intellectual Property Organization. (2008). *WIPO Intellectual Property Handbook* (2nd ed., repr). Geneva: WIPO World Intellectual Property Organization.

Yu, H., & Robinson, D. G. (2012). The New Ambiguity of 'Open Government'. *UCLA Law Review Discourse, 59,* 178–208.

Zanfir, G. (2012). The right to Data portability in the context of the EU data protection reform. *International Data Privacy Law, 2*(3), 149–162. doi:10.1093/idpl/ips009

Zarsky, T. Z. (2003). 'Mine your own business!': Making the case for the implications of the data mining of personal information in the forum of public opinion. *Yale Journal of Law and Technology, 5*(1).

Zhang, K., & Schmidt, Aernout H. J. (2015). Thinking of data protection law's subject matter as a complex adaptive system: A heuristic display. *Computer Law & Security Review, 31*(2), 201–220. doi:10.1016/j.clsr.2015.01.007

Zittrain, J. (2000). What the Publisher Can Teach the Patient: Intellectual Property and Privacy in an Era of Trusted Privication. *Stanford Law Review, 52*.

Zwitter, A. (2014). Big Data ethics. *Big Data & Society, 1*(2). doi:10.1177/2053951714559253